JMS Journées de méthodologie statistique de l'Insee 2018

Simplified variance estimation for multistage sample surveys

Guillaume CHAUVET(*)

(*)Ensai(IRMAR), Campus de Ker Lann, 35170 Bruz chauvet@ensai.fr

Mots-clés. Calage, Enquête ménage, Estimation transversale, non-réponse totale.

Résumé

Les enquêtes auprès des ménages sont souvent sélectionnées selon un plan de sondage à plusieurs degrés. Par exemple, le plan de sondage initial de l'enquête Panel Politique de la Ville [1], réalisée entre 2011 et 2014, peut modulo quelques simplifications être vu comme le résultat d'un plan de sondage à deux degrés. Un échantillon de quartiers est tout d'abord sélectionné, en stratifiant selon le degré d'avancement urbain et avec des probabilités de tirage proportionnelles au nombre de résidences principales. Un échantillon de ménages est ensuite sélectionné dans chaque quartier tiré au premier degré, et tous les individus de 3 ans et plus de ces ménages sont théoriquement enquêtés. Les individus sont suivis pendant quatre vagues d'enquête, avec ajout d'échantillons complémentaires lors des vagues suivantes. Ces ajouts sont réalisés afin de compenser de l'attrition et de permettre de produire des estimations transversales à toutes les vagues d'enquête.

Même dans le cas le plus simple d'une estimation lors de la première vague, l'estimation de variance associée est complexe en raison des différents traitements statistiques. Les poids de sondage des ménages sont ajustés de la non-réponse, en général selon la méthode des groupes homogènes de réponse [2], tout comme les poids individuels. Les poids obtenus sont ensuite calés, généralement de façon simultanée [3], sur des totaux auxiliaires au niveau ménage et au niveau individuel. Pour un plan de sondage à d degrés, la variance d'un estimateur se décompose alors en $d+2$ termes. Les d premiers sont dus aux différents degrés d'échantillonnage. Les deux derniers sont dus à la non-réponse, respectivement de niveau ménage et de niveau individuel.

Dans ce travail, nous étudions les performances d'un estimateur de variance par bootstrap permettant de tenir compte de toutes ces sources d'alea. Cet estimateur ne nécessite pas de produire d'estimateur de variance à l'intérieur des unités primaires, ce qui le rend particulièrement simple d'utilisation. Il conduit généralement à une estimation de variance conservative : il surestime la variance de premier degré, mais estime correctement la variance due aux degrés suivants de tirage. Son utilisation est illustrée sur des exemples.

Cet article est basé sur des travaux réalisés pour le Commissariat Général à l'Egalité des Territoires (CGET) [6], et pour le Household Finance and Consumption Network (HFCN) [7].

Abstract

Multistage sampling designs are commonly used for household surveys. If we wish to perform longitudinal estimations, individuals from the initial sample are followed over time. If we also wish to perform cross-sectional estimations at several times, additional samples are selected at further waves and mixed with the individuals originally selected. Even in the simplest case when estimations are produced at the first time with a single sample, variance estimation is challenging since the different sources of randomness need to be accounted for, along with the needed statistical treatments (correction of unit non-response at the household and at the individual level, correction of item non-response, calibration). In this work, we consider a bootstrap solution which accounts for the features of the sampling and estimation process. This bootstrap solution is usually conservative for the true variance, in the sense that the sampling variance tends to be overestimated. The proposed bootstrap is illustrated with examples.

1 Cross-sectional estimation

We consider estimation at time t for a population U_{hou}^t of households. If y_k^t denotes the value taken by some variable of interest for the household k at time t, we may be interested in the estimation of the total

$$Y_{hou}^t = \sum_{k \in U_{hou}^t} y_k^t. \quad (1)$$

We may also be interested in the associated population U_{ind}^t of individuals. If y_l^t denotes the value taken by some variable of interest for the individual l at time t, the parameter of interest is

$$Y_{ind}^t = \sum_{l \in U_{ind}^t} y_l^t. \quad (2)$$

1.1 Sampling design

We suppose that the sample used for estimation is selected by means of a multistage sampling design. At the first stage of sampling, a population of Primary Sampling Units (PSUs) U_I is defined and partitioned into H strata $U_{I1}, \ldots, U_{IH}$. For example, U_I may be a set of municipalities stratified according to some measure of size. The sample S_I of PSUs is selected in U_I by means of stratified sampling. We note π_{Ii} the inclusion probability of some PSU u_i, and S_{Ih} the sample of n_{Ih} PSUs selected in the stratum U_{Ih}. The design weight of the PSU u_i is

$$d_{Ii} = \frac{1}{\pi_{Ii}}. \quad (3)$$

Note that the sample of PSUs S_I and the first-stage inclusion probabilities π_{Ii} are assumed to be constant over time.

At the second stage and inside any PSU $u_i \in S_I$, a sample of households is selected. We note $\pi_{k|i}^t$ for the conditional probability that the household k is selected into the PSU u_i, and $S_{i,hou}^t$ for the sample of households selected in u_i at time t. The conditional weight of some household k is

$$d_{k|i}^t = \frac{1}{\pi_{k|i}^t} \text{ for any } k \in u_i, \quad (4)$$

and the non-conditional sampling weight of some household k is

$$d_k^t = d_{Ii} \times d_{k|i}^t \text{ for any } k \in u_i. \tag{5}$$

In case of full response of the households, the estimator of Y_{hou}^t is

$$\hat{Y}_{hou}^t = \sum_{h=1}^{H} \sum_{u_i \in S_{Ih}} d_{Ii} \hat{Y}_{i,hou}^t \quad \text{with} \quad \hat{Y}_{i,hou}^t = \sum_{k \in S_{i,hou}^t} d_{k|i}^t y_k^t \tag{6}$$

$$= \sum_{k \in S_{hou}^t} d_k^t y_k^t, \tag{7}$$

with S_{hou}^t the global sample of households. Both equations are of separate interest : equation (6) is useful to obtain a suitable variance estimator for $\hat{Y}_{hou}^t$, whereas equation (7) is simpler for point estimation and makes use of the non-conditional sampling weights only.

We suppose that if a household k is selected at time t, all the individuals within are surveyed. Therefore, the design weight for some individual l at time t is

$$d_l^t = d_k^t \text{ for any } l \in k. \tag{8}$$

In case of full response of the individuals, the estimator of Y_{ind}^t is

$$\hat{Y}_{ind}^t = \sum_{l \in S_{ind}^t} d_l^t y_l^t, \tag{9}$$

with S_{ind}^t the global sample of individuals.

1.2 Treatment of non-response

In practice, the sample S_{hou}^t is prone to unit non-response, which leads to the observation of a sub-sample of respondents $S_{r,hou}^t$ only. We note r_k^t for the response indicator of a household k, and p_k^t for the response probability of the household k. We suppose that the households respond independently of one another. Also, we suppose that unit non-response is handled through the method of Response Homogeneity Groups (RHGs), which is popular in practice. Under this framework, it is assumed that the sample S_{hou}^t may be partitioned into C RHGs denoted as $S_{1,hou}^t, \ldots, S_{C,hou}^t$ such that the response probability p_k^t is constant inside a RHG.

We note p_c^t for the common response probability inside the RHG $S_{c,hou}^t$. It is estimated by

$$\hat{p}_c^t = \frac{\sum_{k \in S_{c,hou}^t} \omega_k^t r_k^t}{\sum_{k \in S_{c,hou}^t} \omega_k^t}, \tag{10}$$

with ω_k^t some weight attached to the household k. One customary choice is $\omega_k^t = 1$, which leads to estimating p_c^t by the response rate inside the RHG. This will be referred to as *unweighted estimated response probabilities*. Another customary choice consists in using the design weights $\omega_k^t = d_k^t$. This will be referred to as *weighted estimated response probabilities*.

Accounting for the estimated response probabilities leads to the weights corrected for non-response

$$d_{rk}^t = d_{Ii} d_{rk|i}^t \quad \text{with} \quad d_{rk|i}^t = \frac{d_{k|i}^t}{\hat{p}_{c(k)}^t}, \tag{11}$$

with $c(k)$ the RHG to which the household k belongs. The estimator of Y^t_{hou} adjusted for non-response is

$$\hat{Y}^t_{r,hou} = \sum_{h=1}^{H} \sum_{u_i \in S_{Ih}} d_{Ii} \hat{Y}^t_{ri,hou} \quad \text{with} \quad \hat{Y}^t_{ri,hou} = \sum_{k \in S^t_{i,hou}} d^t_{rk|i} r^t_k y^t_k \tag{12}$$

$$= \sum_{k \in S^t_{r,hou}} d^t_{rk} y^t_k. \tag{13}$$

Concerning the response probabilities, the unweighted and the weighted estimators are expected to perform similarly in terms of bias. We advocate for the use of weighted probabilities, which leads to a calibration property for the estimator adjusted for non-response. We note

$$\hat{N}^t_{c,hou} \equiv \sum_{k \in S^t_{c,hou}} d^t_k \tag{14}$$

the estimator of the size of a RHG, making use of the design weights d^t_k. Using weighted probabilities enables to match exactly these estimated sizes, in the sense that the calibration equation

$$\sum_{k \in S^t_{c,hou}} \frac{d^t_k r^t_k}{\hat{p}^t_k} = \hat{N}^t_{c,hou}$$

holds true. Consequently, the variance of $\hat{Y}^t_{r,hou}$ is expected to be reduced, as compared to the same estimator using the unweighted estimated response probabilities, if the variables defining the RHGs are partly explanatory for the variable of interest y^t_k.

In practice, the individuals living in the responding households in $S^t_{r,hou}$ are also prone to non-response, though it is expected to be to a smaller extent. This leads to the observation of a sub-sample of respondents $S^t_{r,ind}$ only. We note r^t_l for the response indicator of an individual l, and p^t_l for the response probability of the individual l. We suppose that the individuals respond independently of one another. Also, we suppose that this non-response is handled through the method of RHGs. This leads to the estimated response probabilities $\hat{p}^t_l$, and to the weights adjusted for non-response

$$d^t_{rl} = \frac{d^t_{rk(l)}}{\hat{p}^t_l} \text{ with } k(l) \text{ the household containing } l. \tag{15}$$

The estimator of Y^t_{ind} adjusted for non-response is

$$\hat{Y}^t_{r,ind} = \sum_{l \in S^t_{r,ind}} d^t_{rl} y^t_l. \tag{16}$$

1.3 Calibration

Lastly, the weights adjusted for non-response are calibrated on some auxiliary totals known on the population of households and on some auxiliary totals known on the population of individuals. We note $x^t_{hou,k}$ for the vector of calibration variables at the household level, and $x^t_{ind,l}$ for the vector of calibration variables at the individual level.

We assume that an "integrative" calibration is performed, in the sense that the calibration of the weights d^t_{rk} and of the weights d^t_{rl} is performed jointly on both sets of calibration totals. This

may be done by means of the CALMAR2 software [3], for example. The individual auxiliary variables are first aggregated to the household level, by computing for any $k \in S^t_{r,hou}$

$$x^t_{ind,k} = \sum_{l \in k} \frac{r^t_l x^t_{ind,l}}{\hat{p}^t_l}. \tag{17}$$

The calibration is then performed at the household level on the set of calibration variables

$$x^t_k = \left(\{x^t_{hou,k}\}^\top, \{x^t_{ind,k}\}^\top \right)^\top. \tag{18}$$

For the sample $S^t_{r,hou}$ of households, this leads to the calibrated weights w^t_k and to the calibrated estimator

$$\hat{Y}^t_{cal,hou} = \sum_{k \in S^t_{r,hou}} w^t_k y^t_k. \tag{19}$$

For the sample $S^t_{r,ind}$ of individuals, this leads to the calibrated weights

$$w^t_l = \frac{w^t_{k(l)}}{\hat{p}^t_l}, \tag{20}$$

with $k(l)$ the household containing the individual l, and to the calibrated estimator

$$\hat{Y}^t_{cal,ind} = \sum_{l \in S^t_{r,ind}} w^t_l y^t_l. \tag{21}$$

The sampling and estimation steps are summarized in Figure 1.

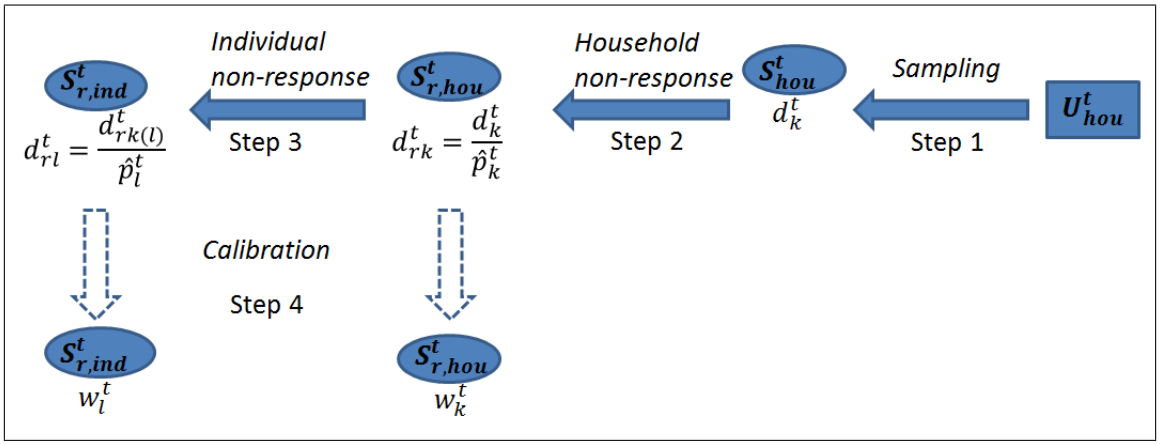

FIGURE 1 – Sampling and estimation steps for a cross-sectional estimation with a single sample

2 Computation of bootstrap weights

2.1 The full response case

In case of full response of the households, the estimator for the population of households is $\hat{Y}^t_{hou}$, which is given in equation (7). A simple variance estimator is obtained by treating the sample

as if the PSUs were selected with replacement. This leads to the variance estimator

$$\hat{V}_{wr}\left(\hat{Y}^t_{hou}\right) = \sum_{h=1}^{H} \frac{n_{Ih}}{n_{Ih}-1} \sum_{u_i \in S_{Ih}} \left(d_{Ii}\hat{Y}^t_{i,hou} - \frac{\sum_{u_j \in S_{Ih}} d_{Ij}\hat{Y}^t_{j,hou}}{n_{Ih}} \right)^2 . \tag{22}$$

This variance estimator would be unbiased if the PSUs were selected with replacement. It can be shown that when the PSUs are selected with a sampling design which is more efficient than with-replacement sampling, this variance estimator is conservative for the true variance. That is, this variance estimator will tend to over-estimate the true variance. Examples of sampling designs which are more efficient than with-replacement sampling include simple random sampling without replacement, in case of sampling with equal probabilities ; and conditional Poisson sampling [8], Sampford sampling [9] and pivotal sampling [10,11], in case of sampling with unequal probabilities.

Bootstrap weights may be obtained as follows. Inside the sample of PSUs S_{Ih} selected in the stratum U_{Ih}, we draw a with-replacement resample S_{Ih*} of $n_{Ih}-1$ PSUs, selected with equal probabilities. For any $u_i \in S_{Ih}$, we take

$$W_{Ii} \;=\; \frac{n_{Ih}}{n_{Ih}-1} \times \text{Number of times the PSU } u_i \text{ is selected in the resample } S^*_{Ih}. \tag{23}$$

The bootstrap version of the estimator $\hat{Y}^t_{hou}$ is

$$\hat{Y}^t_{hou*} \;=\; \sum_{h=1}^{H} \sum_{u_i \in S_{Ih}} W_{Ii} d_{Ii} \hat{Y}^t_{i,hou} \tag{24}$$

$$\;=\; \sum_{k \in S^t_{hou}} d^t_{k*} y^t_k, \tag{25}$$

where

$$d^t_{k*} \;=\; W_{Ii(k)} d^t_k \tag{26}$$

is the bootstrap sampling weight of the household k, with $u_{i(k)}$ the PSU containing k. It can be shown that these bootstrap weights enable to match the with-replacement variance estimator given in (22), in the sense that

$$V^*\left(\hat{Y}^t_{hou*} \middle| S^t_{hou}\right) \;=\; \hat{V}_{wr}\left(\hat{Y}^t_{hou}\right), \tag{27}$$

with $V^*(\cdot)$ the variance under the resampling scheme.

2.2 Adjustment of bootstrap weights for non-response and calibration

In practice, the sample of households is prone to unit non-response (see Section 1.2) which leads to the sub-sample of households $S^t_{r,hou}$. Also, the weights adjusted for non-response are finally calibrated on some auxiliary totals (see Section 1.3). The final estimator is $\hat{Y}^t_{cal,hou}$ given in equation (19).

The computation of the bootstrap weights for the households is described in Algorithm 1. In order to compute bootstrap weights for individuals, the non-response of individuals should in theory be taken into account, since it introduces an additional variability. In practice, it is expected that the non-response of individuals is small to moderate as compared to the non-response of

households. In such case, the additional variability may be safely neglected, and the bootstrap weight for individual l is

$$w_{l*}^t = \frac{w_{k(l)*}^t}{\hat{p}_l^t}, \tag{28}$$

with $k(l)$ the household containing the individual l, and where the calibrated bootstrap weights w_{k*} are obtained as described in Algorithm 1. Note that in the presentation of Algorithm 1, Step 1 (sampling), Step 2 (household non-response) and Step 4 (calibration) refer to the sampling and estimation steps presented in Figure 1. As previously mentioned, the variability associated to Step 3 (individual non-response) is neglected.

Algorithm 1 Computation of bootstrap weights accounting for non-response and calibration in case of multistage sampling

— Step 1 : we account for the sampling design by computing, for any $k \in S_{hou}^t$, the initial bootstrap weight d_{k*}^t given in (26), as described in Section 2.1.

— Step 2 : we account for the non-response of households. We first compute the bootstrap estimated probabilities inside the RHGs

$$\hat{p}_{c*}^t = \frac{\sum_{k \in S_{c,hou}^t} W_{Ii(k)} \omega_k^t r_k^t}{\sum_{k \in S_{c,hou}^t} W_{Ii(k)} \omega_k^t}, \tag{29}$$

with $u_{i(k)}$ the PSU containing the household k. We then compute the bootstrap weights corrected for non-response

$$d_{rk*}^t = \frac{d_{k*}^t}{\hat{p}_{c(k)*}^t}, \tag{30}$$

with $c(k)$ the RHG containing the household k.

— Step 4 : we account for the calibration. For this, the bootstrap weights d_{rk*}^t are calibrated on the estimated totals

$$\hat{X}_{r,hou}^t = \sum_{k \in S_{r,hou}^t} d_{rk}^t x_k^t. \tag{31}$$

This leads to the bootstrap calibrated weights w_{k*}^t.

2.3 Example 1 : computation of the bootstrap weights with two-stage sampling

We now describe a small example to illustrate the construction of bootstrap weights. We consider a population U_{hou}^t of $N_{hou}^t = 100$ households, which are clustered into 7 PSUs. We suppose that a single stratum of PSUs is used, so that we simply note $U_{I1} \equiv U_I$. Among the $N_I = 7$ PSUs in the population, a sample of $n_I = 4$ PSUs is selected with probabilities proportional to the number of households. Suppose that the PSUs u_1, u_2, u_3 and u_4, whose inclusion probabilities are $\pi_{I1} = 0.2$, $\pi_{I2} = 0.4$, $\pi_{I3} = 0.8$ and $\pi_{I4} = 0.8$, are selected.

Inside any of these 4 PSUs, we draw a sample of $n_0 = 3$ households through without-replacement simple random sampling. This two-stage sampling scheme leads to a self-weighted sampling design, where for any household the sampling weight is $d_k^t = \frac{100}{12}$. For example, the PSU u_1 contains

5 households and a sample of 3 households is selected inside, so that $d^t_{k|1} = \frac{5}{3}$ for any $k \in u_1$. On the other hand, we have $d_{I1} = \frac{1}{\pi_{I1}} = 5$ so that from equation (5) $d^t_k = \frac{100}{12}$ for any $k \in u_1$. A summary is given in the left hand-side of Figure 2.

The initial sample of households is $S^t_{hou} = \{A, B, \ldots, L\}$. Among these 12 households, 9 only are surveyed due to non-response. It is accounted for by using the method of Response Homogeneity Groups (RHGs), where the households A, I, K and L form a first RHG, and the other households form the second one. Inside each RHG, weighted estimated response probabilities are used. For example, we obtain in the first RHG

$$\hat{p}^t_1 = \frac{\sum_{k \in S^t_{1,hou}} d^t_k r^t_k}{\sum_{k \in S^t_{1,hou}} d^t_k} = \frac{d^t_A + d^t_I + d^t_K}{d^t_A + d^t_I + d^t_K + d^t_L} = \frac{3}{4}. \tag{32}$$

This is summarized in the right hand-side of Figure 2.

The weights accounting for non-response are obtained by dividing the sampling weights by the estimated response probabilities, which leads to the weights corrected for non-response given in the left-hand side of Figure 3. A final calibration step is then applied, so that the weights enable to match exactly the population size $N^t_{hou} = 100$ and an auxiliary total $X^t_{hou} = 60$. Note that from the sampled values for the variable x^t_k, we have

$$\hat{N}^t_{r,hou} = 100 \quad \text{and} \quad \hat{X}^t_{r,hou} = \frac{160}{3}. \tag{33}$$

This leads to the calibrated weights given in the right-hand side of Figure 3.

The bootstrap is performed by first selecting a resample of $n_I - 1 = 3$ PSUs, with replacement and with equal probabilities, among the original sample of PSUs. In this example, we suppose that the PSU u_1 is selected once, and that the PSU u_4 is selected twice. From equation (23), we therefore obtain $W_{I1} = \frac{4}{3}$ and $W_{I4} = \frac{8}{3}$. The bootstrap sampling weights are obtained from equation (26). For example, the household A has a sampling weight $d^t_k = \frac{100}{12}$ and belongs to the PSU u_1 for which $W_{I1} = \frac{4}{3}$. Therefore, we obtain $d^t_{k*} = \frac{100}{9}$. All the non-zero bootstrap sampling weights are given in the left-hand side on Figure 4.

The bootstrap sampling weights are corrected for non-response in the same way than in the original correction of non-response : using the same RHGs, and weighted estimated probabilities. For example, the households A, K and L belong to the same RHG. The households A and K are respondents, whereas the household L is a non-respondent. The bootstrap weighted estimated response probability inside this RHG is

$$\hat{p}^t_{1*} = \frac{\sum_{k \in S^t_{1,hou}} d^t_{k*} r^t_k}{\sum_{k \in S^t_{1,hou}} d^t_{k*}} = \frac{d^t_{A*} + d^t_{K*}}{d^t_{A*} + d^t_{K*} + d^t_{L*}} = \frac{3}{5}. \tag{34}$$

The bootstrap weights accounting for non-response are obtained by dividing the bootstrap sampling weights by the bootstrap estimated response probabilities. This is summarized in the right hand-side of Figure 4.

A final calibration step is then applied, so that the weights enable to match exactly the estimated population size $\hat{N}^t_{r,hou} = 100$ and the estimated total $\hat{X}^t_{r,hou} = 160/3$. This leads to the bootstrap calibrated weights given in Figure 5.

2.4 Example 2 : computation of the bootstrap weights with a direct sampling of the households

In some cases, the sample of households S^t_{hou} is not selected through multistage sampling, but by direct sampling in the population U^t_{hou}. This is in fact a special case of the set-up presented in Section 1.1, where each Primary Sampling Unit u_i is reduced to a single household. The bootstrap algorithm presented in Section 2.2 may still be applied, but households are resampled rather than PSUs.

To fix ideas, we describe a small example. We consider the same population U^t_{hou} of $N^t_{hou} = 100$ households than in Section 2.3, except that this population is not clustered into PSUs. We suppose without loss of generality that a single stratum of households is used, and that a sample of households S^t_{hou} is selected in U^t_{hou} through simple random sampling of size $n_I = 12$. Therefore, the sampling weight is $d^t_k = \frac{100}{12}$ for any household.

The initial sample of households is $S^t_{hou} = \{A, B, \ldots, L\}$. Among these 12 households, 9 only are surveyed due to non-response. It is accounted for by using the method of Response Homogeneity Groups (RHGs), where the households A, I, K and L form a first RHG, and the other households form the second one. Inside each RHG, weighted estimated response probabilities are used. For example, we obtain in the first RHG

$$\hat{p}^t_1 = \frac{\sum_{k \in S^t_{1,hou}} d^t_k r^t_k}{\sum_{k \in S^t_{1,hou}} d^t_k} = \frac{d^t_A + d^t_I + d^t_K}{d^t_A + d^t_I + d^t_K + d^t_L} = \frac{3}{4}. \tag{35}$$

This is summarized in Figure 6.

The weights accounting for non-response are obtained by dividing the sampling weights by the estimated response probabilities, which leads to the weights d^t_{rk}. A final calibration step is then applied, so that the weights enable to match exactly the population size $N^t_{hou} = 100$ and an auxiliary total $X^t_{hou} = 60$. Note that from the sampled values for the variable x^t_k, we have

$$\hat{N}^t_{r,hou} = 100 \quad \text{and} \quad \hat{X}^t_{r,hou} = \frac{160}{3}. \tag{36}$$

The weights adjusted for non-response and the final calibrated weights are given in Figure 7.

The bootstrap is performed by first selecting a resample of $n_I - 1 = 11$ households, with replacement and with equal probabilities, among the original sample of households. In this example, we suppose that the household A is not selected, that the household B is selected twice, that the household C is selected three times, and so on. The number of times each household is selected in the resample is given in the top part of Figure 8. The bootstrap sampling weights d^t_{k*} are obtained from equation (26). For example, the household C has a sampling weight $d^t_k = \frac{100}{12}$ and has been selected three times in the resample. Therefore, we obtain

$$d^t_{C*} = \frac{12}{11} \times 3 \times d^t_C = \frac{300}{11}. \tag{37}$$

The bootstrap sampling weights are corrected for non-response in the same way than in the original correction of non-response : using the same RHGs, and weighted estimated probabilities. For example, the households I and L belong to the same RHG. The household I is a respondent, whereas the household L is not. The bootstrap weighted estimated response probability inside this RHG is

$$\hat{p}^t_{1*} = \frac{d^t_{I*}}{d^t_{I*} + d^t_{L*}} = \frac{1}{2}. \tag{38}$$

The selection of the resample and the computation of the bootstrap estimated response probabilities is summarized in Figure 8.

The bootstrap weights accounting for non-response are obtained by dividing the bootstrap sampling weights by the bootstrap estimated response probabilities. A final calibration step is then applied, so that the weights enable to match exactly the estimated population size $\hat{N}^t_{r,hou} = 100$ and the estimated total $\hat{X}^t_{r,hou} = 160/3$. This leads to the bootstrap calibrated weights given in Figure 9.

2.5 Bootstrap variance estimation

In this Section, we consider bootstrap variance estimation. Suppose that we are interested in some parameter defined over the population of households, namely

$$\theta^t_{hou} = f\left(Y^t_{hou}\right), \tag{39}$$

where f is a known function, and where $Y^t_{hou} = \sum_{k \in U^t_{hou}} y^t_k$ with y^t_k a p-vector of characteristics observed for the household k. The estimator of θ^t_{hou} is then

$$\hat{\theta}^t_{hou} = f\left(\hat{Y}^t_{cal,hou}\right), \tag{40}$$

where $\hat{Y}^t_{cal,hou}$ is given in equation (19). The bootstrap variance estimator for $\hat{\theta}^t_{hou}$ is obtained as described in Algorithm 2.

Algorithm 2 Computation of the bootstrap variance estimator for an estimation over the population of households

1. Repeat $B = 1\,000$ times the bootstrap procedure described in Algorithm 1, which leads to the resampling weights w^t_{k*} for the households $k \in S^t_{r,hou}$.
2. Note $\hat{Y}^t_{cal,hou*}(b)$ the bootstrap calibrated estimator of the total, obtained by using the bootstrap calibrated weights $w^t_{k*}(b)$ computed at the bootstrap iteration $b = 1, \ldots, B$. Note $\hat{\theta}^t_{hou*}(b)$ for the associated bootstrap estimator of θ^t_{hou}, obtained by plugging $\hat{Y}^t_{cal,hou*}(b)$ into (40).
3. The Bootstrap variance estimator is

$$\hat{V}^B_{boot}\left(\hat{\theta}^t_{hou}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left\{\hat{\theta}^t_{ind*}(b) - \frac{1}{B}\sum_{b'=1}^{B}\hat{\theta}^t_{ind*}(b')\right\}^2. \tag{41}$$

The bootstrap variance estimator for an estimation over the population of individuals is obtained accordingly. Suppose that we are interested in the parameter

$$\theta^t_{ind} = f\left(Y^t_{ind}\right), \tag{42}$$

where f is a known function, and where $Y^t_{ind} = \sum_{k \in U^t_{ind}} y^t_l$ with y^t_l a p-vector of characteristics observed for the individual l. The estimator of θ^t_{ind} is then

$$\hat{\theta}^t_{ind} = f\left(\hat{Y}^t_{cal,ind}\right), \tag{43}$$

where $\hat{Y}^t_{cal,ind}$ is given in equation (21). The bootstrap variance estimator for $\hat{\theta}^t_{ind}$ is obtained as described in Algorithm 3.

Algorithm 3 Computation of the bootstrap variance estimator for an estimation over the population of individuals

1. Repeat $B = 1\,000$ times the bootstrap procedure described in Algorithm 1, which leads to the resampling weights w_{l*}^t for the individuals $l \in S_{r,ind}^t$, see equation (28).
2. Note $\hat{Y}_{cal,ind*}^t(b)$ the bootstrap calibrated estimator of the total, obtained by using the bootstrap calibrated weights $w_{l*}^t(b)$ computed at the bootstrap iteration $b = 1, \ldots, B$. Note $\hat{\theta}_{ind*}^t(b)$ for the associated bootstrap estimator of θ_{ind}^t, obtained by plugging $\hat{Y}_{cal,ind*}^t(b)$ into (43).
3. The Bootstrap variance estimator is

$$\hat{V}_{boot}^B\left(\hat{\theta}_{ind}^t\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left\{\hat{\theta}_{ind*}^t(b) - \frac{1}{B}\sum_{b'=1}^{B}\hat{\theta}_{ind*}^t(b')\right\}^2. \quad (44)$$

References

[1] Sala M., et Chauvet G (2018). "Redresser une enquête longitudinale : le panel politique de la ville", Journées de Méthodologie Statistique.
[2] Juillard H., et Chauvet G (2018). "Variance estimation under monotone non-response for a panel survey", à paraître dans Survey Methodology.
[3] Le Guennec, J., et Sautory, O. (2002). "Une nouvelle version de la macro CALMAR de redressement d'échantillon par calage", Journées de Méthodologie Statistique.
[4] Chauvet G., et Vallée A.-A. (2018). "Consistency of estimators for two-stage sampling", travail en cours.
[5] Kim J.K., Kim, J.J. (2007). "Non-response weighting adjustment using estimated response probability", Canadian Journal of Statistics, vol 35, pp. 501-514.
[6] Chauvet G. (2018). "Rapport méthodologique sur l'enquête Panel Politique de la Ville : estimation de variance". Rapport pour le Commissariat Général à l'Egalité des Territoires.
[7] Chauvet G. (2018). "Methodological Report for the Household Finance and Consumption Network : bootstrap variance estimation for a cross-sectional estimation". Report for the Household Finance and Consumption Network. [8] Hájek, J. (1964). "Asymptotic theory of rejective sampling with varying probabilities from a finite population". Annals of Mathematical Statistics, 35, pp. 1491-1523.
[9] Sampford, M.R. (1967). "On sampling without replacement with unequal probabilities of selection". Biometrika, 54, pp. 499-513.
[10] Deville, J-C., et Tillé, Y. (1998). "Unequal probability sampling without replacement through a splitting method". Biometrika, 85, pp. 89-101.
[11] Chauvet, G. (2012). "On a characterization of ordered pivotal sampling". Bernoulli, 18, pp. 1320-1340.

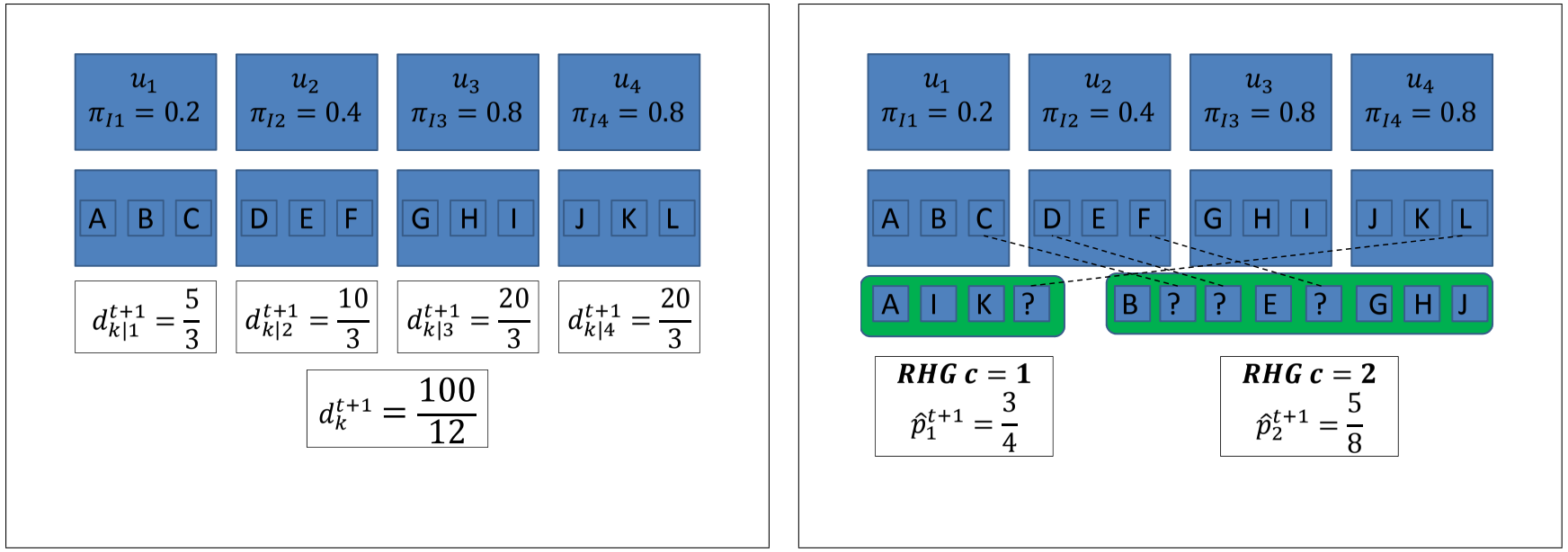

FIGURE 2 – Two-stage sampling : selection of a sample of households (left-hand side) and correction of unit non-response through Response Homogeneity Groups (right-hand side)

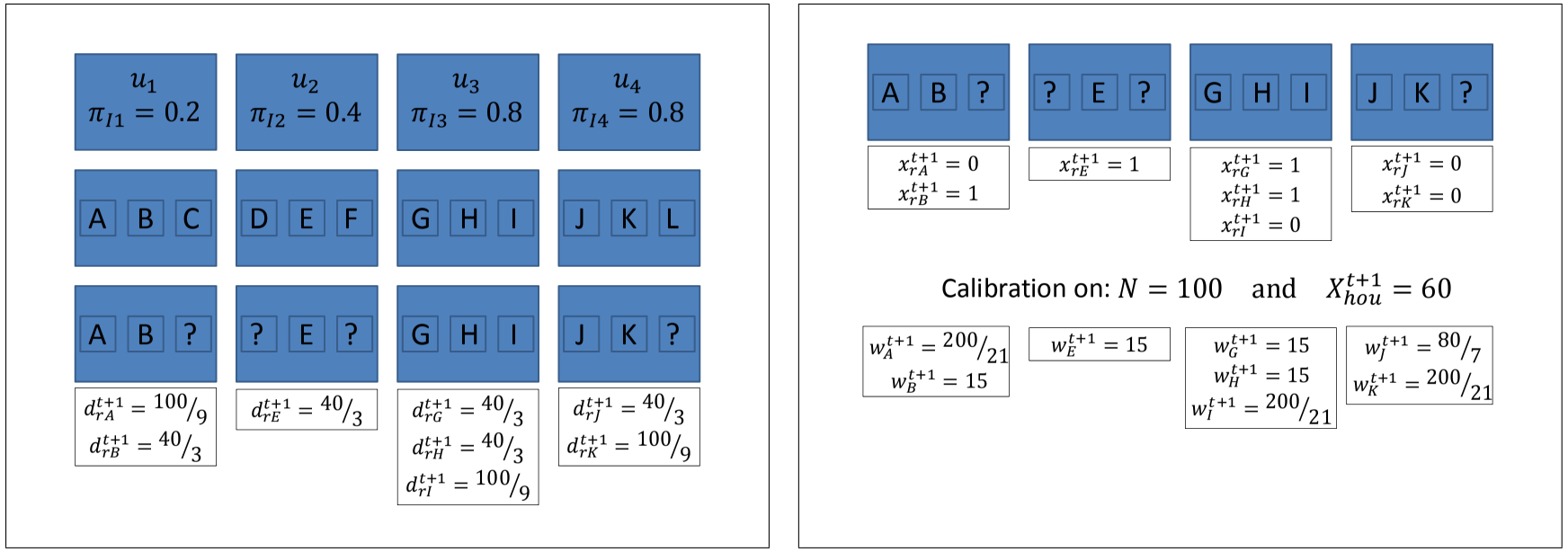

FIGURE 3 – Two-stage sampling : correction of unit non-response of households (left-hand side) and calibration of weights (right-hand side)

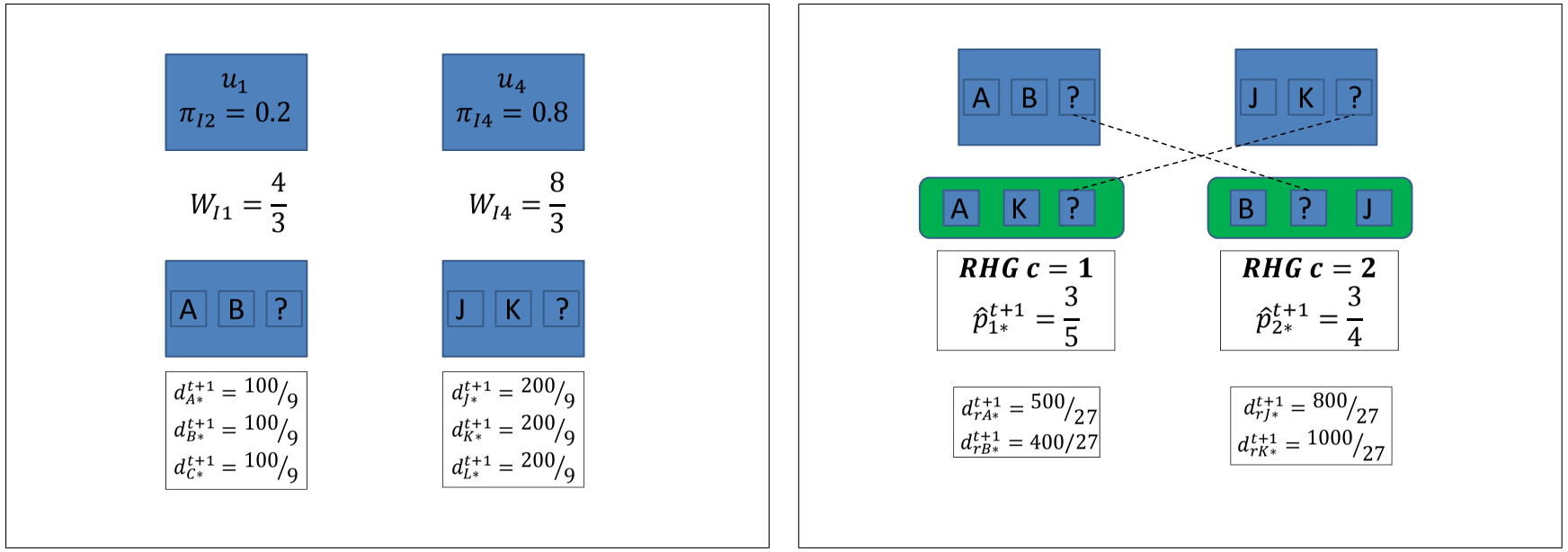

FIGURE 4 – Two-stage sampling : computation of the bootstrap sampling weights (left-hand side) and of the bootstrap weights corrected for non-response (right-hand side)

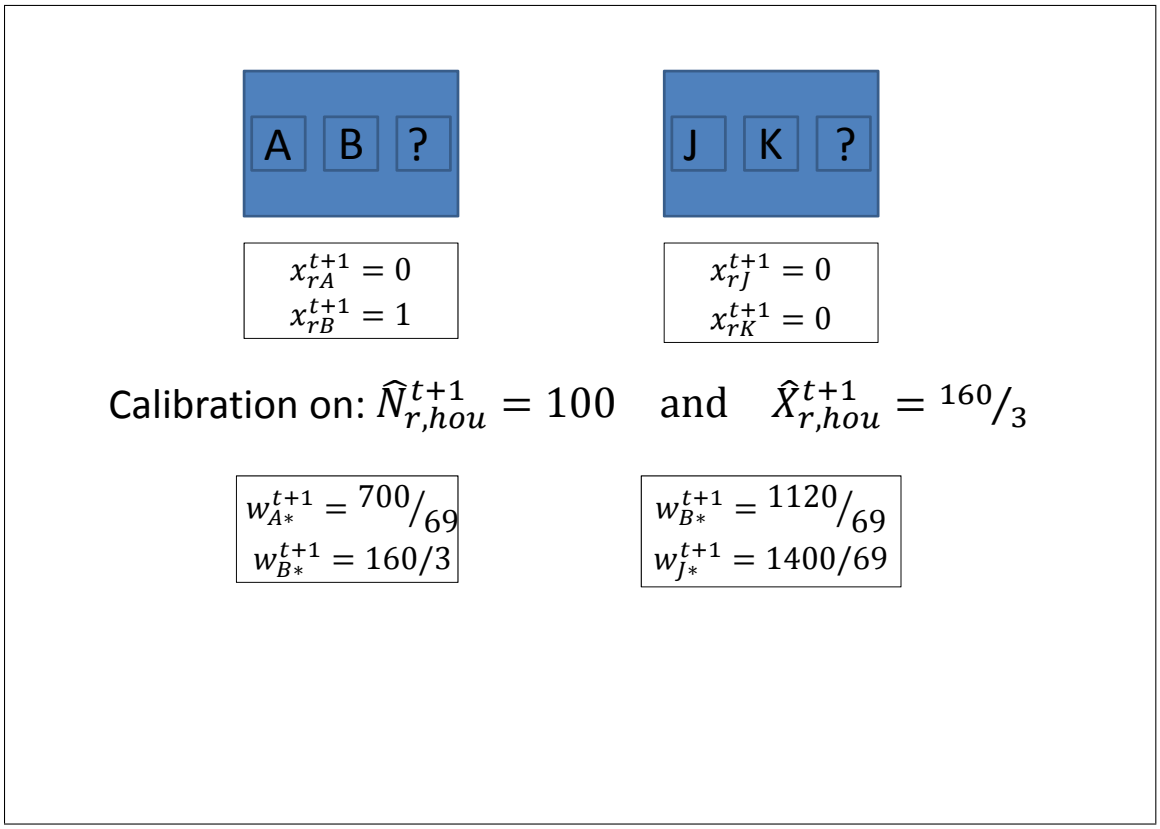

FIGURE 5 – Two-stage sampling : computation of the bootstrap calibrated weights

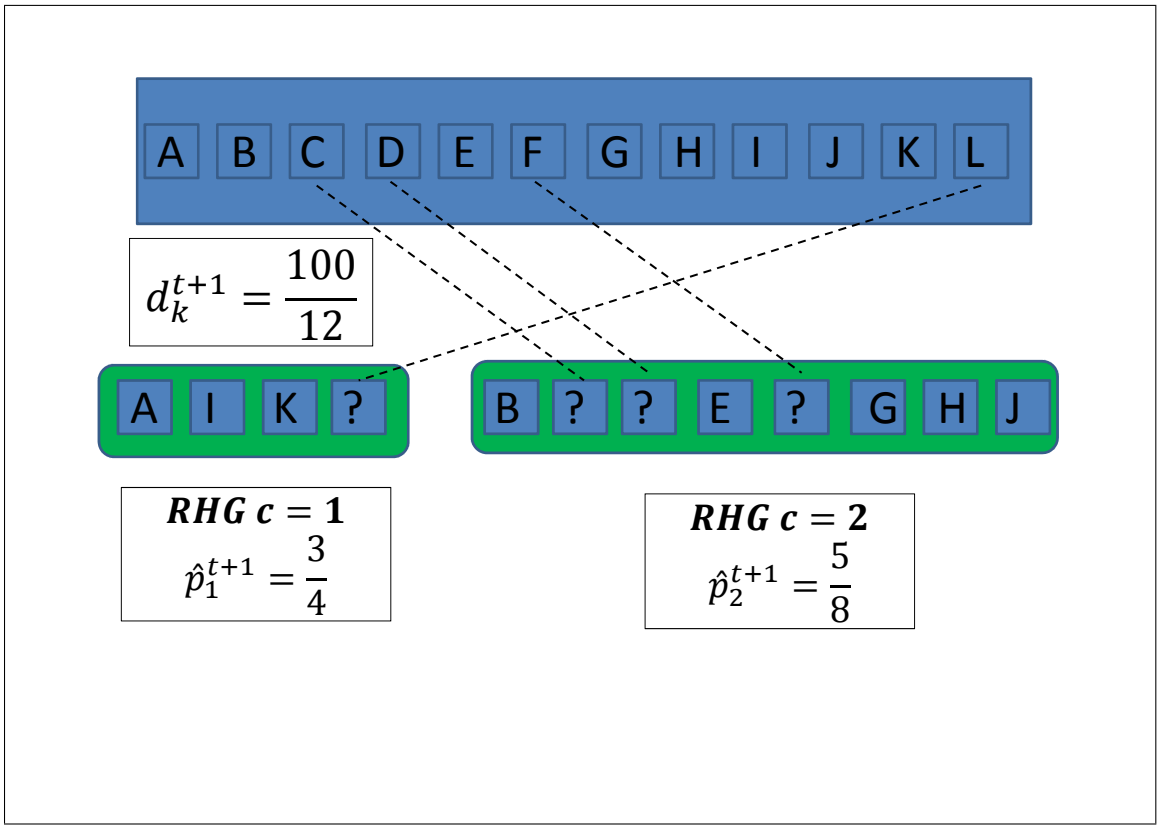

FIGURE 6 – One-stage sampling : selection of a sample of households and correction of unit non-response through Response Homogeneity Groups

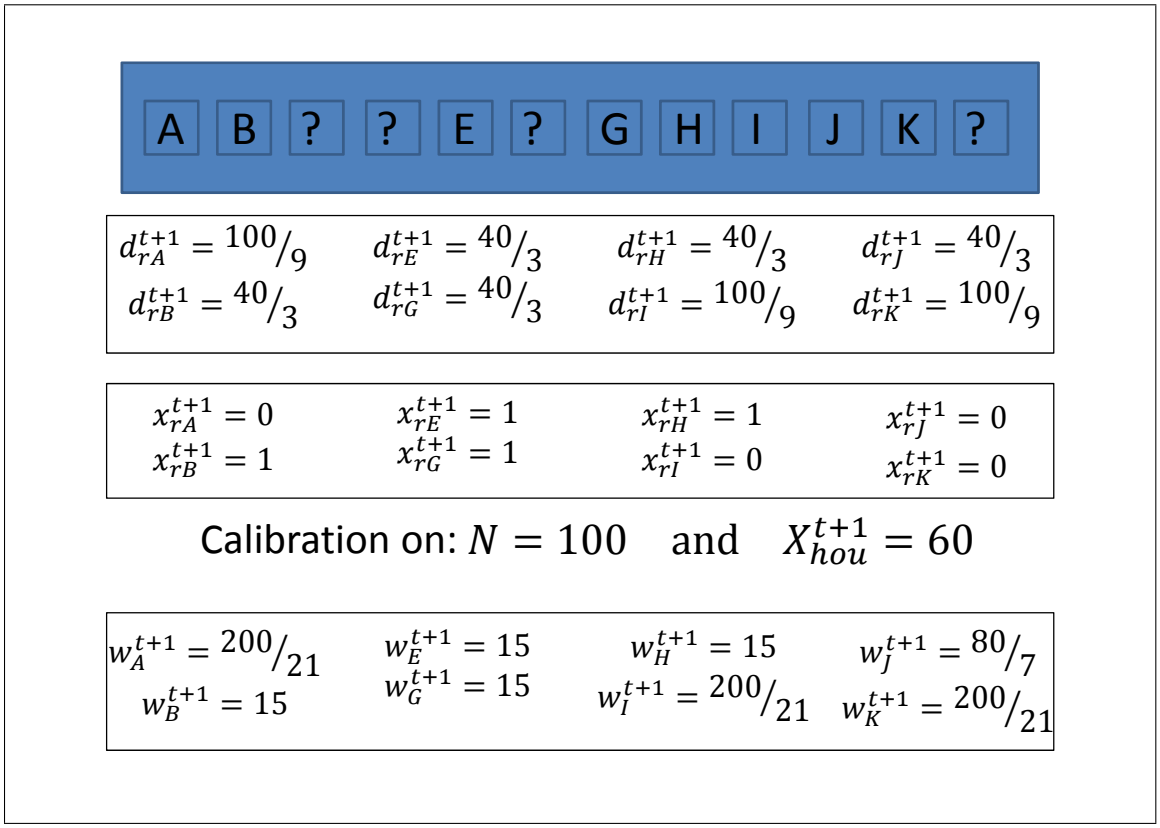

FIGURE 7 – One-stage sampling : correction of unit non-response of households and calibration of weights

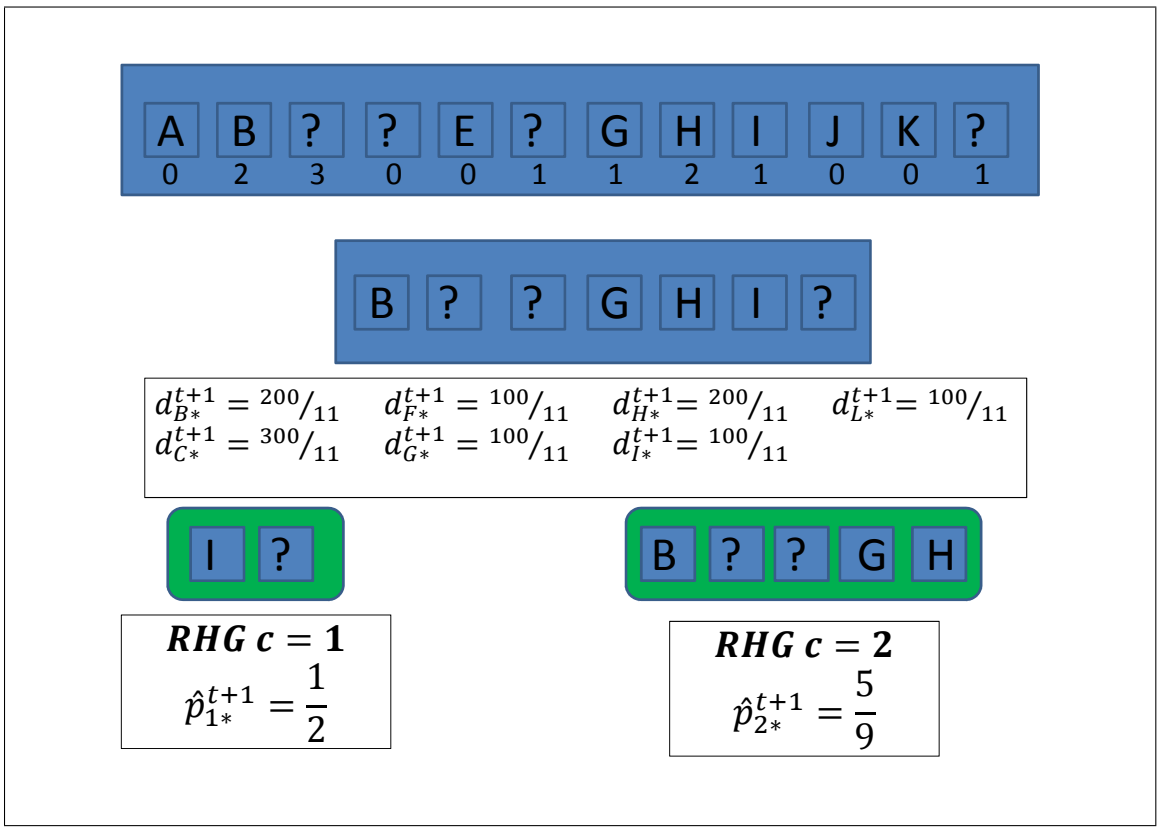

FIGURE 8 – One-stage sampling : Computation of the bootstrap sampling weights and of the bootstrap estimated response probabilities

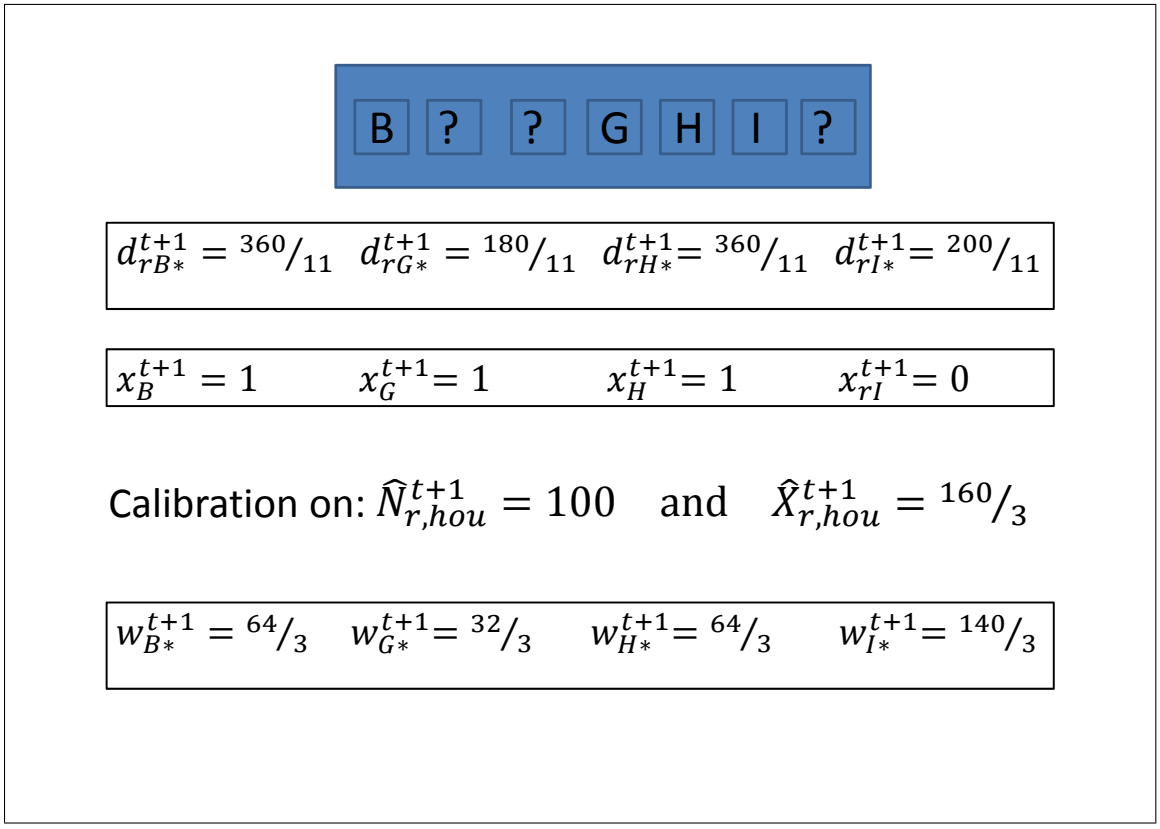

FIGURE 9 – One-stage sampling : Computation of the bootstrap weights adjusted for non-response and of the bootstrap calibrated weights