# Weighting strategy

Jérôme Pasquier, Angéline Chatelan, Murielle Bochud
Institute of Social and Preventive Medicine (IUMSP)
Biopôle 2, Route de la Corniche 10, 1010 Lausanne (Switzerland)

April 2017

menuCH is a survey of the population living in the cantons of Vaud (VD), Geneva (GE), Berne (BE), Neuchatel (NE), Basel-Land (BL), Basel-Stadt (BS), Aargau (AG), Zurich (ZH), St. Gallen (SG), Lucerne (LU), Jura (JU) and Ticino (TI). As in most sampling surveys, subjects do not all have the same probability of being included into the sample. This is why a weighting strategy must be developed and applied to the data. The principle of weighting is about assigning different weights to survey participants based on their probability of inclusion in the sample.

Weighting strategy in menuCH involves three steps:

1. Calculation of the sampling weights;
2. Correction of non-response;
3. Calibration on marginal totals.

These three steps define, for each person who participated in the survey, an extrapolation weight. This latter is used to extrapolate the results of the survey to the target population.

All computations were made using R version 3.3.3 [1] and the survey package version 3.31-5 [2], [3].

## 1. Sampling weights

The sample for the survey menuCH was selected out of the sampling frame SRPH (Stichprobenrahmen und für Personen- Haushaltserhebungen) of the Federal Statistical Office (FSO). The SRPH is the Swiss persons and households registry. The selection of subjects was carried out in five successive waves while the SRPH is updated quarterly. Therefore the sample is composed of five subsamples having been selected in five frames slightly different from each other. The selection was limited to 12 cantons listed in the introduction, thus these cantons represent the target population of the survey.

Within each wave the corresponding sample selection was done according to a stratified sampling design. The sampling frame was divided into 35 strata, defined by the 7 major regions of Switzerland and 5 age categories. All subjects of the same stratum had the same probability of being included in the sample. However, the probability of being included in the sample was different for each stratum. The sampling weights are defined as the inverse inclusion probabilities.

### 1.1 Notation

Let $S_{ij}$ be the set of individuals from the sampling frame belonging to the stratum $i, i \in \{1, 2, \ldots, 35\}$ in the wave $j, j \in \{1, 2, \ldots, 5\}$ and $s_{ij}$ the set of individuals of this stratum who were selected in the sample at this wave.

Let $N_{ij}$ and $n_{ij}$ be the size of the set $S_{ij}$ respectively $s_{ij}$. We also defined

$$\overline{N}_i = \frac{1}{5}\sum_{j=1}^{5} N_{ij}$$

as the mean size of the stratum $i$.

Let

$$s = \bigcup_{i=1}^{35}\bigcup_{j=1}^{5} s_{ij}$$

be the set of individuals selected in the sample. This is a disjoint union and the size of $s$ is given by

$$n = \sum_{i=1}^{35}\sum_{j=1}^{5} n_{ij}.$$

In menuCH $n = 13606$, which corresponds to the number of people who were selected in the sample.

### 1.2 Calculation of sampling weights: first method

The initial weight $w_k$ of an individual $k \in s$ is defined as follows:

$$w_k^{(0)} = \frac{N_{ij}}{n_{ij}}, \quad k \in s_{ij}.$$

To avoid having weights for each wave, a post-stratification by mean size of strata could be carried out. The size of these latter evolves in the range of $-2.8\%$ to $4.4\%$ compared to the first stratum (see Figure 1).

Post-stratified weights $w_k^{(1)}$ are defined as

$$w_k^{(1)} = w_k^{(0)} \frac{\overline{N}_i}{\sum_{j=1}^{5}\sum_{k\in s_{ij}} w_k^{(0)}}, \quad k \in \bigcup_{j=1}^{5} s_{ij},$$

and can be simplified as

$$w_{ijk}^{(1)} = \frac{N_{ij}}{5\, n_{ij}}, \quad k \in s_{ij}.$$

They are used as sampling weights.

However, in menuCH survey, the samples of some waves have not been fully used, therefore some $n_{ij}$ have a very low value ($n_{ij} = 1$ in some cases). This induces a large volatility in sampling weights. That is why this method was not used.

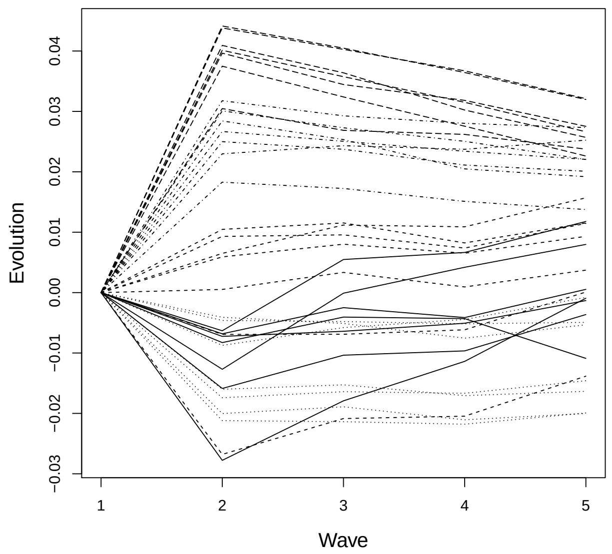


Figure 1: Evolution in size of each stratum in comparison to its size at the first wave.

### 1.3 Calculation of sampling weights: second method

The second method is based on the assumption that the sample was selected in a single wave. In that case weights $w_{ijk}^{(1)}$ are defined as

$$w_k^{(1)} = \frac{\overline{N}_i}{\sum_{j=1}^{5} n_{ij}}, \quad k \in s_{ij}.$$

This method does not take into account that the survey sampling was conducted in several waves but it provides more stable sampling weights. For menuCH we finally decided to use this method.

## 2. Correction for non-response

In view of the substantial erosion in the sample (2086 participants among the 13606 selected people), we tested whether non-response affects the population uniformly or whether certain subgroups respond better than others. The answer can be partly found when comparing the characteristics of non-participants with participants.

### 2.1 Notation

Let $r \subset s$ be set of individuals selected in the sample who participated in the survey and $p_k = Pr(k \in r \mid k \in s)$ the probability of response for the individual $k$.

### 2.2 Non-response model

To determine which variables influenced participation, a logistic regression was performed with the variables available

| | Odds ratio | 2.5% | 97.5% |
|---|---|---|---|
| *age group* | | | |
| 30-39 years | 0.87 | 0.73 | 1.03 |
| 40-49 years | 1.01 | 0.85 | 1.20 |
| 50-64 years | 0.97 | 0.81 | 1.17 |
| >=65 years | 1.10 | 0.88 | 1.36 |
| *sex* | | | |
| Female | 1.19 | 1.08 | 1.31 |
| *marital status* | | | |
| Married | 1.00 | 0.86 | 1.16 |
| Widowed | 0.67 | 0.47 | 0.96 |
| Divorced | 0.90 | 0.74 | 1.10 |
| Others | 1.41 | 0.60 | 3.29 |
| *major region* | | | |
| Midland | 0.97 | 0.83 | 1.13 |
| Northwest Switzerland | 0.99 | 0.84 | 1.17 |
| Zurich | 1.13 | 0.96 | 1.34 |
| Eastern Switzerland | 1.03 | 0.87 | 1.22 |
| Central Switzerland | 1.01 | 0.84 | 1.21 |
| Ticino | 0.83 | 0.70 | 1.00 |
| *nationality* | | | |
| Foreign | 0.41 | 0.36 | 0.47 |
| *household size* | | | |
| 2 people | 1.19 | 1.01 | 1.40 |
| 3 people | 1.19 | 1.00 | 1.43 |
| 4 people | 1.30 | 1.08 | 1.56 |
| 5 or more | 1.21 | 0.99 | 1.49 |

Table 1: Coefficients of the non-response model.

for all people included in the sample. The following variables were considered:

- age group (5 levels: 18-29 years, 30-39 years, 40-49 years, 50-64 years, 65 and over);
- gender (2 levels: male, female);
- marital status (5 levels: single, married, widow, divorced, other);
- major region (7 levels: Lake Geneva region (VD/GE), Midland (BE/NE/JU), Northwest Switzerland (BS/BL/AG), Zurich (ZH), Eastern Switzerland (SG), Central Switzerland (LU), Ticino (TI));
- nationality (2 levels: Swiss, foreign);
- household size (5 levels: 1 person, 2 people, 3 people, 4 people, 5 people or more).

The coefficients of the non-response model are shown in Table 1. It appears that nationality is the main factor associated with non-response.

### 2.3 Non-respondent classes

Predicted response probabilities are not directly used to correct sampling weights but they are needed to define classes

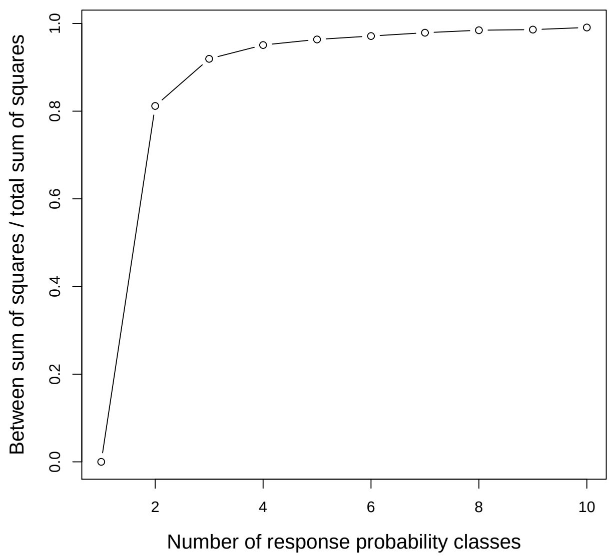


Figure 2: "Between-classes" sum of squares divided by the total sum of squares according to the number of classes.

of non-response using the score method (see [4] p. 33-34). The probabilities are partitioned using k-means clustering. To determine the number of classes, we plotted the "between-classes" sum of squares divided by the total sum of squares according to the number of classes (see Figure 2). We observed that considering three classes can explain most of the variability in the non-response probability. Let $C_1$, $C_2$ et $C_3$ be the three classes obtained using the method of k-means. The mean of the response probability in each of the three classes is respectively $8\%$, $16\%$ et $20\%$ (rounded values).

### 2.4 Weigths after correction for non-response

The response probability for individual $k$ is estimated by

$$\hat{p}_k = \begin{cases} 0.08, & k \in C_1, \\ 0.16, & k \in C_2, \\ 0.20, & k \in C_3, \end{cases}$$

and the weights after correction for non-response are defined as

$$w_k^{(2)} = \frac{w_k^{(1)}}{\hat{p}_k}.$$

## 3. Calibration on marginal totals

The calibration consists in correcting the weights obtained after the first two steps described above to obtain identical distributions to those in the sampling frame for auxiliary variables, which are assumed to correlate with nutrition. For example, if we assume that gender is correlated to nutrition, we will correct weights so that the sum of the calibrated weights for women (respectively men) who participated in the survey matches the number of women (respectively men) from the sampling frame.

### 3.1 Calibration variables

The auxiliary variables used for calibration are the same as those considered in the model of non-response, i.e. age group, gender, marital status, major region, nationality and household size.

### 3.2 Definition of calibration

To meet the goals of the calibration, the calibrated weights $w_k^{(3)}$ must satisfy the following equation:

$$\sum_{k \in r} w_k^{(3)} \boldsymbol{x}_k = \sum_{k \in U} \boldsymbol{x}_k =: \boldsymbol{t}_x,$$

where $r$ is the set of responders, $U$ all people included in the sampling frame and $\boldsymbol{x}_k$ the vector containing the auxiliary information for the individual $k$. Furthermore, the calibrated weights $w_k^{(3)}$ should be as close as possible to original weights $w_k^{(2)}$. They must thus minimize the sum

$$\sum_{k \in r} G(w_k^{(3)}, w_k^{(2)})$$

where $G$ is a distance measure. Here the distance corresponding to the method of raking ratio was chosen. Note that if the auxiliary information must be known for all responders, it is not needed for all individuals of the sampling frame. Only the totals, contained in the vector $\boldsymbol{t}_x$, must be known. The calibration approach and the different distances possible are described in [5].

### 3.3 Average sampling frame

Since the survey was conducted in five waves, five sampling frames are to be considered. For the calibration, a average frame will be considered.

Let

$$N_j = \sum_{i=1}^{35} N_{ij}$$

be the number of individuals contained in sampling frame from wave $j$ and

$$\overline{N} = \frac{1}{5} \sum_{j=1}^{5} N_j$$

the mean number of individuals included in the sampling frame.

Let $\boldsymbol{t}_{x,j}$ be the totals of auxiliary variables for the wave $j$. They are called calibration totals.

The mean totals are defined as

$$\bar{\boldsymbol{t}}_x = \frac{1}{5} \sum_{j=1}^{5} \frac{\overline{N}}{N_j} \boldsymbol{t}_{x,j}$$

and are used then for the calibration. Calibration totals are shown in Table 2.

| Variable | Value | Wave 1 | Wave 2 | Wave 3 | Wave 4 | Wave 5 | Average |
|---|---|---|---|---|---|---|---|
| total | | 4592411 | 4624009 | 4631292 | 4641849 | 4649831 | 4627878.0 |
| sex | Male | 2286093 | 2304406 | 2306852 | 2312210 | 2316157 | 2305141.3 |
| sex | Female | 2306318 | 2319603 | 2324440 | 2329639 | 2333674 | 2322736.7 |
| age group | 18-29 years | 936274 | 922386 | 930743 | 938159 | 944438 | 934406.8 |
| age group | 30-39 years | 878804 | 880842 | 883206 | 887557 | 891303 | 884341.1 |
| age group | 40-49 years | 983592 | 972487 | 973298 | 975383 | 976985 | 976377.8 |
| age group | 50-64 years | 1183817 | 1214433 | 1212588 | 1211564 | 1210167 | 1206494.8 |
| age group | >=65 years | 609924 | 633861 | 631457 | 629186 | 626938 | 626257.5 |
| marital status | Single | 1515765 | 1566597 | 1564609 | 1562891 | 1552541 | 1552446.5 |
| marital status | Married | 2467116 | 2456453 | 2459728 | 2464783 | 2477862 | 2465224.1 |
| marital status | Widowed | 127707 | 120864 | 122753 | 124918 | 126745 | 124601.9 |
| marital status | Divorced | 471168 | 469124 | 472949 | 477709 | 480830 | 474351.0 |
| marital status | Others | 10655 | 10971 | 11253 | 11548 | 11853 | 11254.5 |
| nationality | Swiss | 3399959 | 3416713 | 3416431 | 3418805 | 3420801 | 3414573.4 |
| nationality | Foreign | 1192452 | 1207296 | 1214861 | 1223044 | 1229030 | 1213304.6 |
| household size | 1 person | 808016 | 799198 | 805699 | 810713 | 817490 | 808226.3 |
| household size | 2 people | 1518901 | 1507239 | 1518647 | 1528701 | 1535553 | 1521812.5 |
| household size | 3 people | 828174 | 839788 | 840277 | 841551 | 839822 | 837918.7 |
| household size | 4 people | 860588 | 884100 | 878259 | 874101 | 869964 | 873402.5 |
| household size | 5 or more | 576732 | 593684 | 588410 | 586783 | 587002 | 586518.1 |
| major region | Lake Geneva region | 886689 | 893145 | 896530 | 899533 | 901724 | 895518.1 |
| major region | Midland | 917544 | 923803 | 924494 | 926434 | 926869 | 923831.5 |
| major region | Northwest Switzerland | 819121 | 824709 | 825820 | 827259 | 828811 | 825144.9 |
| major region | Zurich | 1063989 | 1071207 | 1072343 | 1074931 | 1077261 | 1071946.5 |
| major region | Eastern Switzerland | 362648 | 365639 | 365442 | 366095 | 366523 | 365270.4 |
| major region | Central Switzerland | 286919 | 289305 | 289528 | 290021 | 290437 | 289242.1 |
| major region | Ticino | 255501 | 256201 | 257135 | 257576 | 258206 | 256924.5 |
| linguistic region | German | 3272343 | 3295105 | 3298008 | 3304851 | 3309946 | 3296055.4 |
| linguistic region | French | 1064567 | 1072703 | 1076149 | 1079422 | 1081679 | 1074898.0 |
| linguistic region | Italian | 255501 | 256201 | 257135 | 257576 | 258206 | 256924.5 |

Table 2: Totals of sampling frame by wave and average totals.

### 3.4 Alternative calibration

Food consumption was assessed throught two non-consecutive (one month apart) 24-hour dietary recalls. It is known that nutrition is correlated with seasons (spring, summer, autumn, winter) and weekdays (Mo-Th vs Fr-Su). In menuCH recalls were unevenly distributed according to these two factors. It is why we considered to calibrate the weights on seasons and weekdays (in addition to the previous auxiliary variables). We assigned the season for each participant according the mean date between his two recalls. The calibration totals for each season were determined simply by dividing by four the population total of the average sampling frame (4627878). For weekdays we considered three strata: (1) two recalls between Monday and Thursday, (2) two recalls between Friday and Sunday and (3) one recall between Monday and Thursday and one between Friday and Sunday. The calibration totals of these three strata was determined by multiplying the population total of the average sampling frame by $\frac{16}{49}$, $\frac{9}{49}$ and $\frac{24}{49}$ respectively. Note that 28 participants had only one recall. For them, season and weekday strata were determined on the basis of a single date. In addition, one participant had no recall at all. For him, no season and weekday calibrated weight was computed.

### 3.5 Extrapolation weights

The weight obtained after calibration are those used for performing extrapolations to the target population for the variables of interest. A summary of the weights obtained after each stage of the weighting process is presented in Table 3.

## 4. Weigthing for SPADE

To derive usual intakes distributions of foods and nutrients we used Statistical Program for Age-adjusted Dietary Assessment (SPADE). SPADE requires two recalls per participant to assess within participant variance. As previously outlined, 29 participants had only one recall or no recall at all. We thus repeated the non-response and the calibration process considering only the 2057 participants with two recalls as respondents. A summary of the weights thus obtained appears in Table 3.

| | N | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Sum |
|---|---|---|---|---|---|---|---|---|
| Sampling weights | 13606 | 140.1 | 223.0 | 308.3 | 331.6 | 395.6 | 599.5 | 4511830 |
| Non-response weights | 2086 | 704.8 | 1398.0 | 1922.0 | 2175.0 | 2616.0 | 7358.0 | 4536109 |
| Calibrated weights | 2086 | 630.7 | 1419.0 | 1944.0 | 2219.0 | 2619.0 | 9436.0 | 4627878 |
| Season weekday calibrated weights | 2085 | 424.1 | 1246.0 | 1785.0 | 2220.0 | 2638.0 | 18930.0 | 4627878 |
| Non-response weights (2 recalls) | 2057 | 711.9 | 1415.0 | 1949.0 | 2200.0 | 2653.0 | 7575.0 | 4525702 |
| Calibrated weights (2 recalls) | 2057 | 625.4 | 1424.0 | 1988.0 | 2250.0 | 2683.0 | 9991.0 | 4627878 |
| Season weekday calibrated weights (2 recalls) | 2057 | 416.8 | 1249.0 | 1796.0 | 2250.0 | 2645.0 | 21410.0 | 4627878 |

Table 3: Summary of weights for the three steps of the weighting process.

## References


[1] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[2] Thomas Lumley. survey: analysis of complex survey samples, 2016. R package version 3.31-5.

[3] Thomas Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19, 2004. R package verson 2.2.

[4] David Haziza and Jean-François Beaumont. On the construction of imputation classes in surveys. *International Statistical Review*, 75(1):25–43, 2007.

[5] Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992.