# Frequently Asked Questions about Weights in the SHP
# Technical Report

Erika Antal *

March 30, 2017

Besides the different documents explaining which types of weights are available for the Swiss Household Panel (SHP) data-users and how they were constructed, there still remains several questions, when one wants to apply them for analysis. Through a sort of questions-responses this document is intended to help users to correctly apply the weights.

## 1 Frequently asked questions about weights

1. What is the statistical unit of the survey?
   - When doing analysis one has to decide what are the units of his/her research. He/she should also examine if they are the same level as the analysis. In the SHP there are two possible units of survey. The households and the persons.
2. Is the analysis cross-sectional or longitudinal?
   - Basically, if the analysis concerns only one year the analysis is cross-sectional and the weights to be used are the cross sectional weights. If the analysis is based on several years the choice is more complicated.
     - Suppose there is a "reference year", (i.e. a particular year, the population of which we are interested in) and we would like to know how a particular characteristic changes from one time point to an other **in this population**. It is clear that the type of analysis is longitudinal.
     - If the question is something like the change of a particular variable from one time point to an other (e.g. if the average wage of the population of Switzerland changed from one time point to another) the analysis is cross-sectional, even if it concerns several years.

*Swiss Centre of Expertise in the Social Sciences - Bâtiment Géopolis - CH-1015 Lausanne - Switzerland emails: erika.antal@fors.unil.ch

- If the subject of the analysis is the intra-personal change with no particular starting point, cross-sectional weights should be used.

3. What are the differences between the given weights?

   - Generally the differences between the weights variables are indicated in their names. See in the documents here: http://forscenter.ch/en/our-surveys/swiss-household-panel/documentationfaq-2/methods/weighting. We can distinguish:
     - cross sectional vs. longitudinal,
     - inflating to the population size vs. keeping the sample size,
     - concerning the whole combined sample (SHP_I_II_III) vs. only the third sample
     - for the cross sectional weights: individual vs. household,
     - for cross sectional individual weights: children vs. adult,
     - for the longitudinal weights: starting date of the panel in 1999, 2004, 2013 or other year.

4. Which years are concerned in the analysis?

   - As the SHP is a panel survey, it contains data for several years. Weights are computed for each year and are related to this, and only this year. Thus, it is important to clarify not only the population of reference but also the years which are concerned by the analysis. If there is only one year concerned cross sectional weights should be used. If there are several years concerned by the analysis with a reference population, longitudinal weights related to the panel of this particular population are the adequate weights to apply.

5. What is the reference population and why it is important?

   - The reference population is the population for which we would like to make our estimations and inferences. Evidently, when the analysis concerns only one year, the response to this question is quite straightforward. When we would like to follow the population of one particular year (reference year) over years and observe changes or tendencies, the reference population is the population of this particular year. This is the typical case when the longitudinal weights should be used. These longitudinal weights are the most reliable when this year is the first year of a panel (i.e. when the sample was drown). When we pool data from several years and there is no particular year to refer, there is no particular reference population and so longitudinal weights should not be used.

6. Choose the reference year of a longitudinal study...?

- When we would like to know how a particular characteristic changes from one time point to another for a population of a particular year the analysis to perform is clearly longitudinal. The reference year is this year. When we are not interested in following a population of a particular year but rather in
  - a change over a period for a population that changes itself too
  - or changes from one time point to an other, but measured in different persons
  - or comparing something before an event and after an event independently of which year the event happened

  there is no particular year, so no reference year. In these cases, cross-sectional weights should be used and not the longitudinal ones.

7. Longitudinal nature vs. longitudinal weights?

   - There is a lot of incertitude and obscurity about the definition of longitudinal studies. Sometimes any analysis with several years is considered as longitudinal study (e.g. trend studies). Sometimes only those that follow the same units for several time points. What is incontestable is that all the analysis containing observations from several years are longitudinal in nature. But it is far from sure that the adequate weights variables are the longitudinal ones. Longitudinal weights computed in the SHP are adapted for analysis when the interest is in observing changes by tracking the same individuals over several years. When observing the same individuals over time, the changes are less likely to be the result of cultural differences across generations, but rather are in real changes in the behavior of these individuals. Thus, the longitudinal weights should be applied only for these types of research questions.

8. Why the households can not be units of longitudinal analysis?

   - Basically, households can not be the unit of longitudinal analysis. Simply because of their dynamic nature. When the question is about to observe changes over time, we have to make the difference between two types of situations. The one, when the unit (on what the changes are measured) can change itself and in contrast when the unit can not change. Typically, in the case of a household. It is never sure that the composition of the household stays the same from one time point of another. When something is measured on a household at the first time and on the household eventually changed in its composition at the second time, we can not know if the change is due to the change on the value of the measured variable or due to the change on the composition of the household. Imagine a household consisting of a couple with two children in year $t$. The couple gets divorced and the woman moves out of the household with the two children and a new

partner of the man moves in. In year $t + 1$ the former household has a completely different composition and the woman with the two children forms a new household with a new identifier, a household which was nonexistent in year $t$.

9. Why aren't there longitudinal weights for the first year of a sample?

   - At the first year of a panel the reference year of the cross sectional and the longitudinal population are the same. Thus, the cross sectional and the longitudinal weights are the same.

10. "Inflated to the population size" or "keeping the sample size"?

    - It is imperative to choose the right weights for your purpose. As it is written in the User Guide the weights that inflate for the population size are used only to calculate the absolute values of the population. For analysis you should always use the weights that keep the sample size, i.e the ones that have an "S" as the last letter (before the 2 or 3 numbers).

11. The "0", the "-3", the negative weights...?

    - In some cases we can find the value of "0" or "-3" as a weight. They simply mean that the unit is not concern by the analysis using the correspondent weights. The reasons for not being concerned by the analysis can be multiple. The unit may not be in the sample the weights are related to or the unit does not have the required characteristics for having a positive weight value. Other negative values are not possible.

12. Which weights for descriptive statistics?

    - As the difference between the weights that adjust for the population size (...P..) and the one which keep the sample size (...S..) is only a constant proportion $\frac{N}{n}$, both can be applied. For absolute values in general we apply the weights adjusting to the population size.

13. How to do crosstabs with weights in Stata?

    - svyset, clear
      svyset _n [pweight=weights]
      set more off
      svy: tab var1 var2, count obs format(%13.0f)
      svy: tab var1 var2, count format(%13.0f)

14. Constant weights in Stata...?

    - Frequently, Stata commands require constant weights. If only one year is concerned in the analysis, there is only one set of weights, thus they are already constant. When observations for several years

for the same individual are taken in the analysis, there are several sets of weights. In this case, to have constant weights we should take the average of the yearly weights.

15. Type of model - type of weights?

    - There is no one-to-one correspondence between the type of model and the type of weights to use in the analysis. What is important is the reference population and the type of individuals (only original sample members or also other household members) concerned by the analysis. If we are not interested in the evolution (change) of a characteristic of a particular year's population, but rather in a change of this characteristic over a period (in this case the population may - of course - change from year to year) cross sectional weights should be used for **any type** of model. If we are interested in a particular year's population but only in for one year, it is evidently also the cross sectional weights that should be used. In general, longitudinal weights should be used only in the case when the research question is about the **evolution** of **one** $\{particular\ \{year's\ \{population$. In all other cases cross sectional weights are adapted.

16. Pooled cross sectional data + several time points but not the same starting points, pooled regression...?

    - Often, all observations (each individual and each year) are taken for the analysis. In this case there are two aspects that help to choose the adequate weights. Firstly, as each individual has several observations (time points) but with different starting points, there is no particular reference year, or reference population. Secondly, as everyone is taken in the analysis, there are individual (the non-original sample members) for whom there are no longitudinal weights. Consequently, longitudinal weights are not adequate, the cross-sectional ones should be used.

17. Frequency, analytic, probability or importance weights in Stata?

    - In Stata there are four types of weights:
      (a) Frequency weights (**fweight**): They indicates the number of duplicated observations. Unweighted, duplicated data and frequency-weighted data are merely two ways of recording identical information. For example a value 3 means that there are 3 such observations, each identical. They are always integers, important from a data processing perspective, but statistically they are uninteresting.
      (b) Analytic weights (**aweight**): Analytic weights statistically arise in one particular problem: linear regression on data that are themselves observed means. Thus, they are appropriate when

the data in the sample are observed means and typically, the analytic weights are the number of elements that gave rise to the average. The **aweights** are inversely proportional to the variance of the unit's values.

(c) Probability, or sampling weights (**pweight**): They refer to a probability-weighted random sample, and linked to the inclusion probability of a unit in the sample. An observation that had probability 14 of being included in the sample has **pweight** 4. (If there is no adjustment for non-response or other statistical treatments related to weighting system).

(d) Importance weights (**iweight**): Iweights somehow reflects the importance of the observation but have no formal statistical definition. It is rather a catch-all category. They are intended for use by programmers who want to produce a certain computation.

Referring to these descriptions above, with probability random sample - as the SHP - the weights for analysis should be specified as **probability weights**. Note that not every command support every kind of weight. A note below the syntax diagram for a command tell you which weights the command support.

18. What about the within estimator?

- The within estimator is the term of the estimator in a fixed effect model in the context of panel data analysis. Thus the remarks and reflections in point 5 can also be applied here.

19. Which characteristics are used for the computation of the weights?

- During the construction of the weights there are three steps when different variables play a role.
  - First, at the selection stage. The SHP samples were selected by using simple random sample procedures in each of the seven major statistical regions of Switzerland. Thus, the inclusion probability of an individual is determined regionally.
  - Second steps when variables are taken into account for the weights are the adjustments for the non-response at different level (grid, household questionnaire and individual questionnaire). The variables used to adjust for the non-response vary (can vary) depending on the different level and also on the different years. In one year at one of the levels, a variable can be judged as a variable that significantly explains the non-response, when in another year (or the same year, but at a different level) it can be judged insignificant for the non-response. Thus, at these stages, practically each of the variables can be part of the model for constructing the weights.

- The third step is the calibration. Here, the variables taken into account are always the same. They are not based on the data set of the sample but on official statistics of the reference population. These variables are: sex, age, marital status, nationality and region.

20. What about weights of children?

    - Children are persons less than 14 years old and not supposed to fill in the individual questionnaire. They are not original sample members, i.e. they were not selected for the sample, but there is not negligible information about them (in the grid and/or by proxy) that we could use for analysis. For this reason from 2013 onwards cross sectional weights for children were computed.

21. What about the transitional factors?

    - The transitional factors are useful for the development of a "custom made" longitudinal samples. It also allows for the longitudinal weighting of non-original sample members (OSM). In order to simplify the application of weights for longitudinal analysis concerning a sample with an arbitrarily chosen starting date, from Wave 17 onwards the transitional factors were replaced by additional longitudinal weights. To accommodate for these additional weights the SHP has changed the naming conventions for the weights. For more details about the names of the weight variables see: http://forscenter.ch/en/our-surveys/swiss-household-panel/documentationfaq/methods/weighting/.

22. Additional readings about weights.

    - Additional readings about weights can be found here: http://forscenter.ch/en/our-surveys/swiss-household-panel/documentationfaq-2/methods/weighting/.