EXECUTIVE SUMMARY

RECOMMENDATIONS AND PRACTICAL EXAMPLES FOR USING WEIGHTING

February 2008

Sandra PLAZA

Eric GRAF

Correspondence to:

Panel Suisse de Ménages, FORS, Université de Lausanne, Bâtiment Vidy, CH-1015 Lausanne

Tel. +41 (0)21 692 37 30

Fax. +41 (0)21 692 37 35

E-mail: swisspanel@fors.unil.ch

1. Introduction

When analysing survey data, researchers often underestimate the importance of weighting and so they obtain false results. However, using weighting is not sufficient on its own. The customary procedures applied by the various statistical software applications presume that the sample used comes from a straightforward random sampling plan, but this is not often the case.

1.1. Purpose of this document

Why, when and how to use weighting are among the most common questions asked and are fundamental to the analysis of survey data and the estimates relating to them. The answers to these questions can depend on the context and the type of analyses that are to be carried out. There are sometimes solid arguments for deciding whether it is appropriate or not, necessary or not to apply weighting. In the event of a complex survey of limited size, such as the Swiss Household Panel (SHP) or Statistics on Income and Living Conditions (SILC), we think that in the majority of cases the answer is very clear: most of the time it is as necessary as it is useful to apply weighting to compensate for the imperfections in the data sample being used. First of all, the weighting is constructed to take account of differences in the probabilities of inclusion of the units in the sample under consideration. Weighting is essential if there are major differences. High non-response or attrition rates over the various waves of a panel survey are another important reason for applying weighting. Non-response and attrition are often not only important but are also selective and are, for example, more marked for households with incomes at the far ends of the distribution band.

Moreover, the weights have been adjusted to the known totals for the Swiss population (age, sex, nationality, civil status, 7 major geographical regions), which makes the sample more representative. This adjustment reduces any bias in the sample caused by non-response, undercoverage or other distortions, and in some cases also reduces the amount of variance.

It is therefore imperative to use weighting. If the appropriate weighting is not applied, the estimations carried out cannot be considered to be representative of the population under observation. The document (Graf, 2007)[1] provides some advice about choosing the weighting you need to use from all the sets of weights produced.
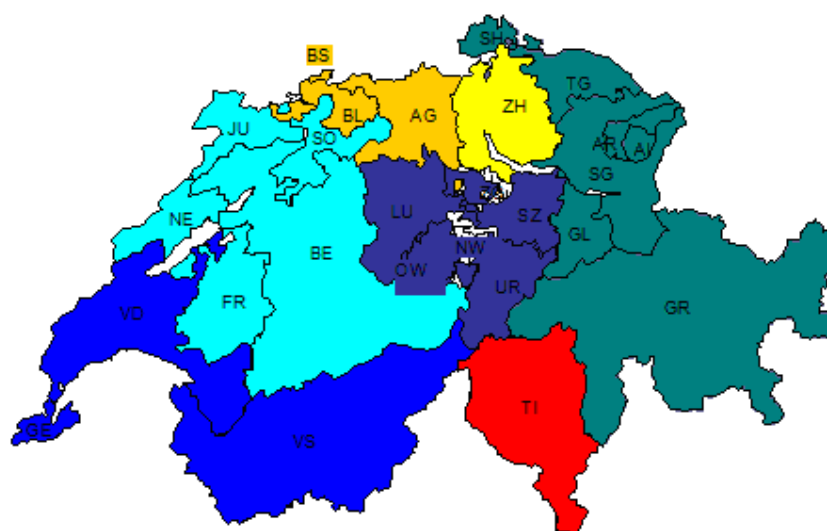
1.2. Survey universe

The SHP covers all individuals living in private households in Switzerland with a (landline or mobile) telephone whose number is in the telephone directory. Individuals living in homes, institutions, collective households or prisons are not covered by the survey. Individuals who cannot be reached by telephone or who have a telephone that is not in the directory are not covered by the survey (undercoverage).

---

[1] Can be downloaded from www.swisspanel.ch/doc/methodology.php?lang=fr&pid=8

It is estimated that in 2004 when the sample for SHP_II was selected, 98.5% of private households were connected to the telephone network. Furthermore, the SRH[2] (the basis for household surveys) covered 93% of these households. In 1999, when the sample for SHP_I was selected, 95% of private households were covered by the SRH.

1.3. Sample

Figure 1: The 7 major geographical regions of Switzerland (NUTS II regions)

The samples are stratified by major region, in proportion to the number of SRH telephone numbers per stratum, without any regional oversampling. In theory, one telephone number equals one household, which is roughly true[3]. We calculated the following proportions:

Table 1: Proportions calculated for the samples in SHP and SILC surveys.

STRATA	SHP_II & SILC (2004, 2005)	SHP_I (1999)
Lake Geneva region (VD, VS, GE)	18.22%	19.13%
Mittelland (BE, FR, SO, NE, JU)	22.92%	22.40%
North-west Switzerland	13.86%	13.18%
Zurich (ZH)	18.22%	17.30%
Eastern Switzerland (GL, SH, AR, AI, SG, GR, TG)	13.70%	14.48%
Central Switzerland (LU, UR, SZ, OW, NW, ZG)	8.75%	8.75%
Ticino (TI)	4.33%	4.76%
Total	100%	100%

---

[2] **The SRH is a registry for household surveys at the Federal Statistics Office and is the basis for the SHP samples in 1999 and 2004.**

[3] **There can sometimes be duplicated data in the SRH (e.g. two different telephone numbers for the same household), or telephone numbers for businesses (2%) or second homes (<1%).**

The SILC and SHP_II surveys have exactly the same sampling plan. It is almost identical to the 1999 plan for SHP_I, except that the percentages that define the proportions per stratum came from censuses in 2000 (for SILC, SHP_II) and in 1999 (for SHP_I).

1.4. Erosion / attrition

Table 2: Number of individuals/households interviewed in the SHP survey. Relative attrition is given by the number of cases interviewed in year (n) over the number of cases in year (n–1).

Year	SHPI	*Individuals* Relative attrition	Absol. attrition	SHPI	*Households* Relative attrition	Absol. attrition	SHPII	*Individuals* Relative attrition	Absol. attrition	SHPII	*Households* Relative attrition	Absol. attrition
1999	7799	100%	100%	5074	100%	100%						
2000	7071	91%	91%	4425	87%	87%						
2001	6593	93%	85%	4139	94%	82%						
2002	5691	86%	73%	3581	87%	71%						
2003	5220	92%	67%	3227	90%	64%						
2004	4406	84%	56%	2831	88%	56%	3648	100%	100%	2538	100%	100%
2005	3885	88%	50%	2457	87%	48%	2648	73%	73%	1909	75%	75%
2006	**4093**	**105%**	**52%**	**2534**	**103%**	**50%**	**2570**	97%	70%	**1682**	88%	66%

As can be seen from the above table, the attrition (or erosion) of the samples from one wave to the next, i.e. the loss of respondents (individuals and households), is a very important issue. In 2005, in its 7th wave, SHP_I only had half the number of respondents it had in 1999.
Major efforts are made, firstly to limit the number of non-responses, but also to improve the models that attempt to correct it in the weighting. But despite that, as attrition is not uniform, some of its characteristics are not taken into account in the models. We know, for example, that for variables that do not correlate or do not correlate much to the variables used to model non-response and in the adjustments of the weighting, it does not change the results much whether weights are used or not. Yet even in those cases, using weights does enable variance calculations that are more correct and realistic (cf. next section).

1.5. Variance calculations, tests, confidence intervals, etc.

The concepts *variance*, *coefficient of variance* and *confidence interval* provide an indication of the amplitude of the sampling variability. However, we remind users that the estimations of these elements[4] provided by the conventional procedures of most common statistics software packages (SAS, SPSS, Stata, MS Excel, etc.) are wrong when the data analysed come from complex surveys such as the SHP or SILC. These procedures assume among other things that straightforward random sampling was carried out. This therefore means that they do not take into account the special characteristics of the survey plan of SHP/SILC such as stratification, nor the corrections made for non-response, nor the post-stratification (raking ratio). Even if most of these procedures enable weighting to be used in producing estimations, they do not take proper account of all the specificities of our surveys when calculating estimations of variance, which is an essential element of most statistical tests.
In fact, there is a tendency to underestimate that variance and consequently the confidence intervals of estimates and the significance thresholds in the tests.

---

[4] We are thinking of elements such as variances, variance coefficients, but also the results of hypothesis testing (e.g. P values accompanying statistics or Pearson's statistics).

The procedures PROC SURVEYFREQ, PROC SURVEYMEANS, PROC SURVEYREG, PROC SURVEYLOGISTIC, etc. in the SAS software enable users to calculate variance using the Taylor series linearization and provide results (at least for SHP and SILC) that are very similar to those obtained using resampling methods that are heavier to implement, such as bootstrapping or jackknifing. Consequently we recommend that SAS users use those procedures that certainly provide results that are more correct than the standard procedures (i.e. PROC FREQ, PROC MEANS, PROC REG, PROC LOGISTIC, etc.).

The STATA software enables users to make the same measurements by adding the 'svyset' and 'svy:' instructions to the conventional syntax. They are applied in exactly the same context and for the same reasons as the PROC SURVEY, etc. in SAS. Some authors (Siller, 2005) have already found that they achieved exactly the same results as with SAS in this way. The analyses that can be carried out with the PROC SURVEY (etc.) procedures in SAS can also be replicated in SPSS using its COMPLEX SAMPLE module. It is therefore possible for almost all users, regardless of the software they are using, to calculate variances using a method adapted to analysing the data from surveys such as SHP and SILC.

1.6. Longitudinal or cross-sectional

Let us just remind ourselves here that cross-sectional weights always refer to the current year, both for households and for individuals, whereas longitudinal weights (individuals) always extrapolate to the population resident in Switzerland in 1999 for SHP_I, and to the population resident in Switzerland in 2004 for SHP_II. The combined longitudinal weightings for SHP_I and SHP_II also refer to the population resident in Switzerland in 2004.

All cross-sectional weights provided extrapolate to the population living in private households in Switzerland in the year of the current wave.

2. Procedures in SAS and SPSS

In both SAS and SPSS software applications there are two types of procedures, those called traditional (it is supposed that the sample is taken from a straightforward random sampling plan) and those procedures that take account of the complexity of the survey plan (SURVEY procedures in SAS and COMPLEX SAMPLE in SPSS). The document 'Recommendations and practical examples for using weighting' together with the help function in SAS and SPSS, provide users with the details and reasons for this[5].

2.1. Elements to be taken into account when applying the SURVEY procedure in SAS and COMPLEX SAMPLE procedure in SPSS

Before using these two types of procedures, users must define various points. To use them correctly it is essential to find out the sampling method used in the survey. In the case of SHP and SILC we have to take stratified straightforward random samples (cf. section 1.3). It is necessary to know the stratifying variable used and the size of each of the strata. In our case,

---

[5] SAS online documentation can be found by following this link:
http://support.sas.com/onlinedoc/913/docMainpage.jsp?_topic=statug.hlp/intro_index.htm. For further information about the SURVEY procedures, go to 'Search' and click on the 'The Survey Procedures' link.
For further information about the COMPLEX SAMPLE procedures in SPSS, go to:
www.spss.com/PDFs/SCS16SPChr.pdf or in the Help section of the software, go to 'Summary' and click on the 'Complex Samples Option' link.

we stratify according to the seven major regions of Switzerland. Depending on whether we are using the data from SHP_I, SHP_II, SILC_I or SILC_II, the size of these strata is not always the same, because the various samples were not taken at the same time. It is also important to choose the right weight. The choice is based firstly on the type of analyses you want to make (longitudinal or cross-sectional, cf. section 1.6); secondly depending on whether you are using individuals or households; and thirdly whether you use the data from SHP_I, SHP_II, SILC_I or SILC_II separately or combined.

When using the COMPLEX SAMPLE procedures in SPSS it is essential to define a subsidiary file called 'plan'[6]. This file contains all the information about the survey plan for that sample and the variable to use for the weightings. In SAS all this information is entered directly into the syntax.

3. Example

This is an example that illustrates the impact and importance of using weights. The analysis relates to the variable *i05htyn* (net annual income of the household) and we wanted to find out the average annual income of a household. In this example we use the data from SHP_I and SHP_II combined and the weighting variable used is variable *wh05t1s* (cross-sectional weight household for SGP_I and SHP_II combined, in which the size of the 2005 sample remains unchanged). We use the procedures PROC MEANS and PROC SURVEYMEANS in SAS and the procedure DESCRIPTIVES in the COMPLEX SAMPLE module in SPSS.

For the analysis with the variable *i05htyn*, we take into account those households whose weight is higher than zero. The sample we use is comprised of 3,697 observations. Table 3 shows the results obtained.

---

[6] The SPSS help function provides the necessary information for using the COMPLEX SAMPLE procedures. To see how to create a plan, go to 'Summary' in the Help function and select 'Sampling for a Complex Design', then click on the 'Show me' link.
Users can find further details in paragraph 2.3.2 and in section 3.1 as well as in Annexes II and IV of the document 'Recommendations and practical examples for using weighting'

Table 3: Summary table of the results obtained with variable i05htyn *for all households.*

Procedure and options used		Mean	stderr	LCL	UCL
SAS	PROC MEANS	96885.394	1142.752	94644.908	99125.879
	PROC MEANS WEIGHT: wh05t1s	91259.119	1104.625	89093.384	93424.854
	PROC SURVEYMEANS STRATA WEIGHT: wh05t1s	91259.119	1167.811	88969.500	93548.738
SPSS	DESCRIPTIVES in the module COMPLEX SAMPLE	91259.119	1167.811	88969.500	93548.738

N : 3697 / Min : 5000 / Max : 1800000

Source: PSM_I and PSM_II

The descriptive analysis of the variable *i05htyn* (net annual income of the household) shows that using weighting has a huge influence on the results. We can see a change in the estimated average, and the confidence intervals for the average estimated with weights and without weights do not correspond.

This indicates that for the analysis of this variable, it is important whether we use weights or not. We could reach different conclusions on account of the differences in the way the average is estimated.

For procedures PROC MEANS with weights and PROC SURVEYMEANS we see that the estimated average is the same. The difference (slight increase in standard intervals with the use of the PROC SURVEYMEANS procedure) comes from the way in which the accuracy is estimated. Moreover, procedure DESCRIPTIVES in the COMPLEX SAMPLE module in SPSS gives exactly the same results as procedure PROC SURVEYMEANS.

4. Conclusion

When analysing data from surveys such as the Swiss Household Panel (SHP) or the Statistics on Income and Living Conditions (SILC) it is necessary to use suitable procedures. The example given shows the impact and the importance of using weighting but also how important it is to take into account the complexity of the survey plan to avoid incorrect results. Furthermore, it is possible to carry out analyses in which the variance calculation is adequate with both SAS and SPSS applications.

It is possible that in other examples the weighting might have little impact but even if the influence of the weighting on the estimator is minor, it should still be used as the variance estimation is more accurate. In those cases where major differences appear, it is of course essential to use weights.

Besides the essential use of weights, as the surveys examined are not carried out according to a straightforward random sampling plan, the customary procedures implemented in the statistical software applications do not provide an adequate calculation of variances.

This is why we recommend the use of SURVEY procedures in SAS. They provide results that are more accurate than those provided by customary procedures. This is because these procedures do not underestimate that variance, or underestimate it less, than in customary procedures. They enable users to take into account all the specific details of the survey.

The SURVEY procedures provide results (at least for SHP and SILC surveys) that are very similar to those obtained using resampling methods, such as bootstrapping or jackknifing.

We should not forget that the SPSS software provides the same facilities as SAS. The SUDAAN and STATA software applications also enable users to make the same measurements, as some authors have shown (Siller, 2005). It is now therefore possible, regardless of the software being used, to calculate variances using a method adapted to analysing the data from surveys such as SHP and SILC and the examples show that we obtain the same results. We should also note that attention should be paid to the way in which SAS and SPSS handle missing values, as they differ somewhat.

To conclude, it is important to use the weights as soon as possible, regardless of the type of analyses to be carried out. Furthermore, it is possible to simply make variance estimates that are more correct and realistic when carrying out analyses on surveys such as SHP and SILC, by using the SURVEY procedure in SAS or COMPLEX SAMPLE procedure in SPSS.