# Technical Note No. 4

## How to Use the Sampling Weights

**HOW TO USE THE SAMPLING WEIGHTS**[1]

**Background:**

The Yemen National Social Protection and Monitoring Survey (NSPMS) consists of a sample of households that were tracked over a twelve-month period. Four rounds of interviews were conducted.

**The sampling weights:**

Since each sampled unit represents a different fraction of the population, we need to weigh observations when carrying out a statistical analysis. In the dataset we have four cross-sectional weight variables: (i) `weight`, (ii) `adj_weight_w2`, (iii) `adj_weight_w3`, and (iv) `adj_weight_w4`. They apply respectively to rounds 1, 2, 3 and 4, and they should be used when carrying out a cross-sectional analysis of a specific round.

In addition to these variables, there are three other longitudinal weight variables: (i) `adj_weight_w1ew2`, (ii) `adj_weight_w1ew2ew3`, and (iii) `adj_weight_w1ew2ew3ew4`. They are suitable for a longitudinal analysis of a set of rounds; (i) is suitable for an analysis including all households interviewed in rounds 1 and 2, (ii) for an analysis including all households interviewed in rounds 1, 2, and 3, and (iii) for an analysis including all households interviewed in all four rounds. This way, once the researcher defines the rounds to be analysed, the weights work with a balanced sample of households for these rounds to be analysed by the researcher.

An appropriate statistical analysis of the data in the NSPMS also requires that we take into consideration the complex survey design of the sample. The NSPMS follows a two-phase sampling design. In the first phase, enumeration areas (clusters) are taken with unequal probabilities from each governorate (stratum). These clusters are our primary sampling units. In phase two, a stratified simple random sample of households is selected from each enumeration area taken in the first phase. In order to take this design into account when using a statistical analysis software, we need to explicitly declare the survey design and configure the software in order to consider it when computing statistics of interest.

**Examples:**

*Individual level variable*

Here we present an example of how to declare the survey design and how to use it to compute, in STATA, the mean of the `absent` indicator for round 1.

The variable specifying the clusters (enumeration areas) is `f15` and the variable specifying the strata (governorate) is `f5`. For example, if we wanted to do a cross-sectional analysis of all

[1] For a detailed report on the calculation of the weights and the statistical procedures related to the sampling design please refer to "Technical Report on the Cross-Sectional Sampling for the Yemen National Social Protection and Monitoring Survey (Vieira and Ferraz, 2013)" and "Technical Report on the Sampling Design for the Yemen National Social Protection and Monitoring Survey (Vieira and Ferraz, 2013)".

households interviewed in round 1 (regardless of whether they were interviewed in subsequent rounds), we would declare this by stating:

```
svyset f15 [pw = weight], strata(f5)
```

In the above command line, we declare that `f15` is our primary sampling unit, `f5` contains the strata, and that we want to use `weight` as a probability weight in our analysis.

Now, to compute the mean of `absent` taking into account the survey design, we combine the command `mean` with the prefix `svy:` as follows:

```
svy: mean absent if f2 == 1
```

The prefix `svy:` tells the command `mean` to take into account the survey design, including `weight` as a probability weight `[pw = weight]` in the declaration of the survey design. Since `weight` applies to the first round, we restricted our estimation to the cross-sectional observations of the first round (i.e., for `f2` equal 1).

In the above example, we computed the mean of `absent` in round 1 for individuals in households that were interviewed in that round, regardless of whether they were followed in other round(s). If we wanted instead to compute the mean of the indicator (in round 1) for individuals in households that were interviewed in all four rounds, we would input the following:

```
svyset f15 [pw = adj_weight_w1ew2ew3ew4], strata(f5)
svy: mean absent if f2 == 1
```

Please note that once we set the survey design together with the weight and we intend to continue using it, we do not need to set it again each time; it is enough to put the `svy:` prefix in front of the command (whenever STATA allows it).

*Household level variable*

The sample units in the NSPMS are the households, but information is collected for each individual living in those households. Hence, if we want to compute statistics based on information that concerns the household, we need to consider each information for a household only once.

Suppose, for example, that we want to know the proportion of households whose main source of light in the first cross-section is electric light from the public grid. The variable containing this information is `electricity1`; it has the value of `1` when this is the main source and `0` otherwise. Assuming that we have not yet set the survey design (and the weight), we would do this as follows:

```
svyset f15 [pw = weight], strata(f5)
svy: mean electricity1 if f2 == 1 & line_no == 1
```

By imposing the restriction `line_no == 1`, where `line_no` identifies an individual within a household, we end up selecting only one individual within the household, and thus we do not consider each household more than once.