Technical Note No. 2

Note on Cross-sectional Weights

Technical Report on the Cross-sectional Sampling Weights for the Yemen National Social Protection and Monitoring Survey

Updated version

Cristiano Ferraz, PhD
Departamento de Estatística
Universidade Federal de Pernambuco, Brazil

Marcel de Toledo Vieira, PhD
Departamento de Estatística
Universidade Federal de Juiz de Fora, Brazil

Recife, October 2014.

1. Introduction

This report describes the statistical procedures adopted for generating the cross-sectional sampling weights for the Yemen National Social Protection and Monitoring Survey (NSPMS). The NSPMS has the Yemeni resident population (excluding non-household communities such as refugees, nomads and internally displaced persons, hotels, dormitories, prisons and hospitals) as its target population. Expected to last for 12 consecutive months, the NSPMS is a longitudinal household survey that aims to provide parameter's estimates quarterly, and to accommodate the Social Welfare Fund (SWF) program impact assessment. The first-quarter survey data is currently available, providing the necessary information for the implementation of the techniques here described.

This report is organised in four further sections. Section 2 describes the main aspects of the NSPMS sample design that guided the sampling weights composition. Section 3 presents the building steps of the sampling weights, and Section 4 gives a general guidance for non-response treatment. Section 5 includes some concluding remarks.

2. The NSPMS sample design

The NSPMS follows a two phase sampling design. In phase one, a stratified cluster sampling design with unequal selection probabilities is taken, where Enumeration Areas (EAs) are considered as the primary sampling units (clusters) selected within each governorate (stratum). In phase two, a stratified simple random sample of households is selected from each EA selected at the first phase. The second phase stratification is based on screening information raised at the first phase sample, and comprises three groups: one treatment and two control groups. Twelve households were sampled from each EA, as presented in Table 1. Thirty EAs are selected from each of the 21 governorates, providing a total sample size of 7560 households.

Table 1. *Second phase strata groups and sample size allocation*

Stratum	Description	Sample allocation
i)	Treatment	5
ii)	Control 1	5
iii)	Control 2	2

Further detailed descriptions on the NSPMS sample design can be found in Vieira and Ferraz (2013).

3. Sampling weights

Cross-sectional sampling weights described in this report reflect not only the NSPMS sampling design but also the application of adjustment terms for dealing with unit non-response cases found at the first wave (i.e. first quarter) data collection process. Retaining notation used by Vieira and Ferraz (2013), let the following quantities be defined:

- π_{hi} is the first phase inclusion probability of EA i within governorate h ;
- $\pi_{g\,j|hi}$ is the second phase conditional inclusion probability of household j within group g given the selected EA i within governorate h ;
- is the inclusion probability of household j within group g , at EA i of
  $\pi_{hig\,j}=\pi_{hi}\pi_{g\,j|hi}$. governorate h .

The starting point for building the sampling weights is the Horvitz-Thompson (1952) estimator for a population total, given by

$$\hat{Y}=\sum_{h}\sum_{i}\sum_{g}\sum_{j} d_{higj}\, y_{higj} \quad ,$$

where

- $d_{higj}=1/\pi_{h\,i\,g\,j}$ is the basic sampling-design-weight for household j within group g , at EA i of governorate h ; and
- y_{higj} is the value of the interest variable y for household j within group g , at EA i of governorate h .

In this simpler form, the basic sampling weights d_{hijg} reflects only the main aspects of sampling design, i.e the inverse of selection probabilities. For instance, since the first phase sampling scheme considers selection with probability proportional to a poverty measure, the estimation process assigns higher sampling weights on those EA with lower poverty levels, and lower sampling weights on those with higher poverty levels.

The first wave data revealed that, as expected in any survey, unit non-response cases were observed. Such cases were assumed to be generated by a missing at random (MAR) mechanism (Rubin, 1976), where the missing pattern depends only on the stratification sampling design variables (i.e. governorate and group).

In order to cope with this problem, a weighting adjustment procedure was adopted to correct the basic sampling-design-weight.

Let $\hat{\varphi}_{hig}=\frac{m_{h(r)}}{m_h}\frac{n_{g/hi(r)}}{n_{g/hi}}$ be the estimated propensity score for responses within group g at EA i whithin governorate h. In this expression, m_h is defined as the number of EAs selected in governorate h, $m_{h(r)}$ is the number of EAs that were actually surveyed, $n_{g/hi}$ is the sample size in the classification stratum *g* within EA *i* of governorate *h*, and $n_{g/hi(r)}$ is the number of households that actually responded the questionnaire in the classification stratum *g* within EA *i* of governorate *h.*

Then:

- $\tilde{w}_{higj}=\frac{d_{higj}}{\hat{\varphi}_{hig}}$ is the sample weight adjusted for unit nonresponse at governorate h, within EA *i* and group g.

The household adjusted sample weight $\tilde{w}$ is still subject to the effect of some households with extreme weight values, that could cause variance inflation. In order to investigate this undesired effect in the households non-response adjusted sampling weights, the adoption of a weight trimming procedure was evaluated following methodology discussed by Potter (1988, Section 2). The trimming method is based upon evaluations of the effect of trimming on the estimates of variables of interest means and sampling variances. Two variables of interest were investigated: total income and total of rural population. After the application of such methodology, the indicated weight trimming value for controlling total income was the 94th percentile of the original non-response adjusted sampling weights. Let $q_{0.94}$ represent the 94th percentile of the original non-response adjusted sampling weights, r_{higj} be the adjustment term for non-extreme original weights, and r'_{higj} be the adjustment term for the extreme original weights. Therefore, the suggested final sampling weight, w_{higj} would be expressed as

- $$w_{higj}=\begin{cases}\tilde{w}_{higj}r_{higj}, & if \quad \tilde{w}_{higj}<q_{0.94}\\ q_{0.94}r'_{higj} & if \quad \tilde{w}_{higj}\geq q_{0.94}\end{cases}$$

However, when analysing results for controlling the total of rural population, no trimming adjustment was recommended. Since both variables are of interest, the final sampling weights do not receive any further adjusting procedure, keeping its expression as introduced before. Hence,

$$w_{higj}=\tilde{w}_{higj}$$

4. Item non-response remarks

Although the cross-sectional sampling weight introduced in Section 3 takes into account unit non-responses for the first wave data, item non-responses eventually found in the survey data set should be handled with additional care.

Imputation of missing item values have the advantage of keeping the information collected for the responded ones. Several imputation techniques are available in the literature. Hot deck techniques consist of taking one of the observed values of the sample as the imputed one. In its simpler form, a partition of the survey data set is determined cross-classifying observations into two or more variables that are available and define the characteristics of the missing data. Within each cell of the partition, a donor is randomly taken to furnish the responded value as the imputed information, when necessary.

When using imputation to cope with item non-response, it is advisable to keep an indicator variable for the imputed values, as they are artificially generated. This information can be useful for estimating appropriate variances and keeping track of the original data set.

5. Concluding remarks

In this report, information of the NSPMS cross-sectional sampling weights building process, and the relevance to appropriately address item non-response problems were provided. These information are useful for the release of official estimates based on the first wave survey data. Improvement of such estimates may be achieved by incorporating available auxiliary information into the estimators functional form, for instance, by further adjusting the cross-sectional sampling weights. The efficiency of these procedures depend on the quality of the relationship between the auxiliary and response variables, reason for which auxiliary information must be carefully chosen by the analysts.

The Yemeni NSPMS can also produce estimates for changes over time, based on longitudinal sampling weights. Details on the calculation of such sampling weights are introduced at Vieira and Ferraz (2014).

A final remark should be made on the unbiasedness of the statistical estimation process based on the NSPMS survey. Since a probability sample was selected, based on a measurable sampling design, the adoption of a Horvitz-Thompson estimator for totals, means and proportions, ensures unbiasedness of the estimates and their respective variances. Furthermore, under the assumption of a missing at random non-response mechanism, the corrections applied to the basic sampling weights also let the Horvitz-Thompson estimators based on the corrected weights retain the same desired statistical properties (Isaki and Fuller, 1982; Vieira, 2009).

6. References

Horvitz, D. G. and Thompson, D. J. (1952) A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, Vol. 47, No. 260, pp. 663-685.

Potter, F. (1988) Survey of Procedures to Control Extreme Sampling Weights. *Procedures of the Survey Research Methods Section*, American Statistical Association.

Rubin, D. B. (1976) Inference and Missing Data, *Biometrika*, 63 pp. 581-590

Vieira, M. D. T. and Ferraz, C. (2014) *The Technical Report on the Sampling Design for the Yemen National Social Protection and Monitoring Survey*. Updated version. Technical Report. Brasília, International Policy Centre for Inclusive Growth.

Vieira, M. D. T. and Ferraz, C. (2014) *The Technical Report on the Longitudinal Sampling Weights for the Yemen National Social Protection and Monitoring Survey*. Updated version. Technical Report. Brasília, International Policy Centre for Inclusive Growth.

Isaki, C. T. and Fuller, W. A. (1982) Survey Design Under the Regression Superpopulation Model. Journal of the American Statistical Association. Vol. 77, n. 377, 89-96.

Vieira, M. D. T. (2009). Analysis of Longitudinal Survey Data. Saarbrücken: VDM Verlag.