Sample weights for Sauti za Wananchi

Weights are calculated for individual respondents and households in survey data in order for the weighted survey sample to represent the population of interest as closely as possible. The Sauti za Wananchi (SzW) survey data are no exception and with the public release of the household baseline data individual respondent and household weights are provided. This document describes aspects of the calculation of the SzW design weights.

Generally, there are three types of survey weights (Kalton and Flores-Cervantes, 2003)[1]:

A. **Design or base weights**: these weights are used to compensate for unequal sample selection probabilities across sample units and are the inverse of the sample selection probabilities. These weights are included with the SzW baseline PAPI household data.
B. **Non-Response weights**: these weights are required to compensate for unequal non-response rates across groups of respondents. These weights are calculated for each CATI round. We provide some attention to non-response weights in section B. below but these weights are not provided with the data; researchers using the data will need to calculate and add these when using data from the call rounds.
C. **Post-stratification weights**: these are additional weights to adjust sample characteristics to conform to known population values, for example sex-age distributions or population totals. Calculation of these weights is also left to data users.

This note describes calculation of weights for the SzW baseline PAPI household data. The baseline data also contain community, school and health facility data. Weighting for these datasets is discussed briefly in section D.

A. Design weights

This section sets out the approach for calculating sample design weights for the Sauti baseline household data. In the baseline sampling a three stage process has been used, entailing:

1. Selection of enumeration areas (EAs) – undertaken by PPPS using Census frame 2012.
2. Selection of households – by random sampling from EA household list
3. Selection of respondent – random selection among the household members aged 18 or above

Selection of EAs was done from two strata, urban and rural, with the number of EAs sampled for each in proportion to the EA stratum population. We therefore do not have to correct for the relative strata sizes.

Calculation of design weights

[1] Kalton, Graham and Ismael Flores-Cervantes, Weighting Methods, *Journal of Official Statistics*, Vol.19, No.2, 2003. pp. 81–97.

The first stage in the calculation of design weights reflects differences in the probability of inclusion of each respondent in the sample. Generally speaking, this probability is the product of the selection probability at each of the three stages: probability of EA selection times probability of household selection times probability of adult respondent selection. For a respondent *i* living in household *j* located in EA *k* (in each stratum *l*) the generic formula for sample selection probability is

1) $$P_{ijkl} = \frac{1}{NA_j} \times \frac{\propto}{NH_k} \times \frac{\beta}{NE_l}$$

where α gives the number households sampled in EA k; β gives the number of EAs sampled from strata l; NA_j is the total number of adults in household *j;* NH_k is the total number of households in strata *k*; and NE_l is the number of EAs in stratum *l* (but the last term in the equation is equal for urban and rural strata by design). The design weights are calculated as the inverse of the selection probabilities. We discuss the components of equation (1).

EA selection term (3rd term on RHS): this is constant across the two strata, since we selected the same population proportion of urban and rural EAs from the 2012 Census frame.

HH selection term (2nd term on RHS): here we have to deal with two issues. First, recall that two types of respondents were selected by teams in the field: (a) ten "main sample" (or original) households were selected from among those households where network reception at the dwelling was "sufficient/good" and (b) two replacement respondents from households where network reception was sufficient/good *and* a mobile phone was present at the time of the baseline survey (these households did not receive a phone). This approach has the following implications for the design weights.

For **households owning a phone at the baseline**, the second term on the RHS of equation (1) representing the household sample selection probability is, omitting subscripts,

2) $$P(hh, phone) = p_1 + (1 - p_1)p_2 = \propto \frac{1}{NH_k^*} + (1 - \propto \frac{1}{NH_k^*}) \propto_R \frac{1}{NH_k^p}$$

In equation (2), p_1 is the probability to be selected for the main sample, p_2 the probability of selection into the reserve sample; α gives the number households sampled in EA k; NH^* is the total number of households with network reception in the EA; α_R is the number of reserve households and NH^p is the total number of households with a phone at baseline in the EA as noted in the listing form. Equation 2 shows that households owning a phone have a (slightly) larger selection probability as they can be selected as both reserves and as main respondents.

For **households *without* a phone at baseline**, the only selection possibility is in the first round when main respondents are selected. For these households, selection probability is simply

3) $$P(hh, no\ phone) = p_1 = \propto \frac{1}{NH_k^*}$$

According to equations (2) and (3), the design weights will differ within EAs between households on the basis of phone ownership at baseline. The Sauti design weights that are provided with the data have been calculated using these equations.

A second issue is that the EAs were selected with PPPS from the census EA frame, which means the EA *total* number of households (from the 2012 Census) has already been incorporated at EA sample selection stage; in other words, this term is implicit in the sample. However, the sample was drawn not from all households but from those around whose dwelling network reception was sufficient. In other words, we need to correct the expressions to incorporate this.

B. Non-response weights

Note that after the pen and paper interviews (PAPI), once the panel respondents are being interviewed by call center enumerators on their mobile, the sample composition will change through "response selection": not all respondents will respond during call rounds. Moreover, original respondents drop out and are replaced by reserve households to form a new respondent base. These issues require reweighting the sample to mitigate bias due to these non-random events.

Non-response weighting requires a theory of non-response to select sample characteristics and variables needed to estimate weights. In the Sauti sample a number of variables appear relevant. First, household wealth is likely to be important as a proxy of household welfare. It is not a priori clear whether the correlation with response probability is positive or negative: wealthy respondents could be too busy to respond but so could poor respondents; moreover, poor respondents have reported to sell their phone which ended their role in Sauti. Second, location (urban/rural) seems important: rural respondents are more likely to live in areas with network reception difficulties/fluctuations. Third, being a farmer may mean moving often to a field outside network reception area. Respondent characteristics like age, gender, education level could also be correlated with response probability. Finally, a factor that is directly related to the sample design (weight) is phone availability in the household at baseline and will be used in the response modeling as well.

We use logistic regression weighting (Kalton and Flores-Cervantes, 2003) to estimate weights. Using design weights only as well as in combination with non-response weights, we investigate effects on estimates of mean age and conclude that

- Design weights make a small difference in the mean age estimate (less than 0.3%)
- Adding non-response weights makes a slightly larger difference in the mean age estimate (relative to design weights only) of about 1%
- Using decile averaged non-response weights rather than individual response weights makes a very small difference in the mean age estimate of about 0.05%. The "advantage" of a lower standard error estimate is equally small, about 0.015%. For this reason we typically do not pursue this weight adjustment.
- The overall conclusion is to use both design weights and non-response weights in all estimates (except for estimates based on PAPI baseline data only, where only design weights are needed).

C. Post-stratification

The SzW EA sampling was done using the 2012 Census frame and the now published Census results in principal could be used for post-stratification. However, note that the population of interest for SzW is adult respondents from households living in areas with “sufficient” network coverage. We leave it to data users to decide on this type of correction.

A final note is on aggregate population statistics: the design weights are calculated for the full set of households sampled, 2000 main and 400 reserve households. The sum of these weights approximates the population (of mainland Tanzania). When aggregate statistics are needed but not all reserves are used the weights should be adjusted to reflect this.

D. Weights for EA level data

The SzW baseline PAPI data also contain community, school and health facility data. These are EA level data since the survey was designed to administer one community, school and health facility questionnaire for each EA. For these datasets the weighting is, in principal, unnecessary since sampling of Enumeration Areas (EAs) was done with Probability Proportional to Population Size (PPPS). In other words, EAs with larger populations had a larger probability of selection and thus design weights are not needed.

However, not all 200 EAs returned information for these questionnaires, as follows:

- Community data: 193 observations
- School data: 156 observations
- Health facility data: 114 observations

To start with the community data, the 7 missing observations (out of 200) are non-response. The enumerators were tasked to administer the community questionnaire to the community leadership: village or mtaa leader. The reason for non-response given by the data collection teams was that the EA representative could not be found at the time of the EA visit. It is not easy to judge whether this non-response is random but the non-response is limited.

The absence of EAs in the school and health data is not necessarily “non-response”. The data collection teams were instructed to collect data from schools and health facilities within or close to the EA, so non-representation includes true absence of schools and health facilities in the EA (vicinity). However, in some cases non-response was due to non-cooperation by school and health facility staff. Preliminary checks did not reveal a substantial significant correlation between most observable EA characteristics and “non-response”. However, school questionnaire non-response is significantly higher in urban EAs.

In conclusion, we recommend that the EA level data are used without design weights. Researchers are, however, recommended to explore whether non-response weights apply.