UNHCR
The UN Refugee Agency

Forced Displacement Survey

SAMPLING NOTE

V 1

Forced Displacement Survey

PAKISTAN

2024

Contents

Target populations / sampling universe

The Forced Displacement Survey Programme, globally, aims to collect data on forcibly displaced populations and their hosts.

In Pakistan, the FDS collected data from a nationally representative sample of the refugee population as well as a sample of their hosts, i.e. nationals who live in close proximity to the refugee population and whose lives are affected by their presence. These constitute the two primary target populations of the survey; each is described below in turn.

Refugees

The definition of the refugee population is based on UNHCR's determination of their status. Hence all of the individuals registered by either UNHCR or host governments that have undergone a successful process of status determination are considered eligible for the purpose of the Forced Displacement survey. This definition aligns with UNHCR official statistics emanating from registration databases and reported in products such as the Global Trends Report and Refugee Data Finder. This definition will also provide a reliable base for comparison of the Forced Displacement data across contexts.

It must be noted that the UNHCR and host government status determination activities may not capture all of the refugees in a given host country due to various reasons. The Forced Displacement survey will continue exploring methods to capture these groups as experimental samples in order to provide a better understanding of the total potential refugee population in the host country. With this ambition, the survey aims to address possible under-coverage of the refugee population in UNHCR registration systems as well as provide evidence on how to improve registration through operational means.

FDS collects household level information for refugees. It is therefore important to also define what constitutes a refugee household for the purpose of data collection as well as analysis. To-date there is no fully agreed upon definition of the refugee household. The definition of the refugee household for the FDS is one where the head of household or his/her spouse is a refugee.

The main sampling frame for the refugee population globally is thus the UNHCR registration data system proGres. In Pakistan the registration data system, DRIVE, is maintained by the government of Pakistan and managed jointly with UNHCR.

The refugee population in Pakistan is almost entirely from Afghanistan. Pakistan is not signatory to the 1951 Convention on the Status of Refugees, or its 1967 protocol, but has generously supported Afghan refugees for decades. Thus, Pakistan hosts around 3 million Afghanis, however with variable legal status and designations. Afghanis in Pakistan consist of 3 legal groups:

- 1,306,599 Proof of Registration (PoR) card holders

- 840,000 Afghan Citizen Card (ACC) holders
- 775,000 undocumented

The PoR card holders constitute the primary refugee population targeted in the Forced Displacement Survey. Detailed registration information on this group is available from a high-quality recent DRIVE registration exercise.

These populations are divided between formal refugee villages and those living at large among the host population, either in rural or urban regions. The refugee villages are primarily concentrated in Khyber Pakhtunkhwa (KPK) province as well as in Balochistan with one village situated in Punjab province. Refugee villages are formally managed and supported by UNHCR jointly with the government Commissionerate for Afghan Refugees (CAR) and are typically associated with a host community village, typically the adjacent or attached to the village, with free movement of people and goods between the village and settlement. The villages are in most cases clearly distinct from the host villages, however, with regular and uniform dwellings, and fully populated by refugee populations. Outside these formal settings, refugees live in both rural and urban areas with quite different characteristics.

The main sample for the refugees in Pakistan is spread across the following strata geographically

- Khyber Pakhtunkhwa (KPK) province,
- Balochistan province
- Punjab province and
- urban locations in Islamabad and Karachi.

Refugees in North, East and Adamaoua are predominantly refugees from Central African Republic (CAR) and they live both in refugee settlements as well as in the communities at large. This sampling stratum was therefore further divided into three distinct allocation strata:

- Refugees living in settlements,
- Refugees living out of settlements in larger townships
- Refugees living out of settlements in rural villages

Refugees in all three provinces living in rural settings reside both in refugee villages as well as live at large among the host population. This sampling strata for the three provinces (excluding urban stratum) are therefore further divided into two distinct allocation strata:

- Refugees living in refugee villages,
- Refugees living at large among the host population out of refugee villages

This table summarizes the refugees by province and indicates proportion living within refugee villages.

Of the PoR card holders, 54% are children, 22% women, 31% with disabilities. 32% live in refugee villages, while 68% live alongside hosts in urban and rural areas.

The survey (sampling universe) will be representative of the entire PoR card holder population in the country.

Table 1. Refugee population in Pakistan

Province	Count	Proportion	Proportion living in RV
Khyber Pakhtunkhwa	687,877	52.6 %	49.4 %
Balochistan	315,733	24.2 %	17.4 %
Punjab	186,901	14.3 %	7.4 %
Sindh	72,428	5.5 %	0.0 %
Islamabad	39,342	3.0 %	0.0 %
P.A.K.	4,318	0.3 %	0.0 %
Total	1,306,599	100.0 %	31.0 %

Non-PoR Card Holding Afghans (NPCA)

Less is known of Afghan Citizen Card (ACC) holders and undocumented Afghans. In the FDS, these two groups are considered together as "non-PoR card holding Afghans" (NPCA) and are being captured in an experimental sample.

ACC holders were previously undocumented Afghans residing in Pakistan. The Government of Pakistan issued ACCs to this population after an agreement was reached with the Afghan government in 2017-2018. The formal validity period of ACCs has since expired. Unlike PoR cards, ACC cards do not entitle bearers to particular rights, and do not require establishing internationally accepted criteria for refugee status.

IOM reports that 20% of ACC holders live in refugee villages. Beyond this, we do not have reliable data on the location or geographical distribution of NPCA population.

NPCA Afghan refugees are not included into the main survey sample.

Host community

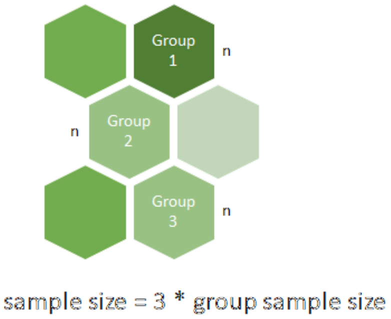
There is no standardised and agreed-upon definition of the refugee host population, globally or in Pakistan. FDS globally and in Pakistan particularly aims to contribute to the discourse on the operationalization of a definition of host communities that can be used both in the survey as well as in operational contexts. Therefore, FDS aims to capture the living conditions of Pakistan nationals both for comparison with refugees and to understand the consequences of living in proximity to the refugees, especially for those in proximity to refugee villages.

In accordance with operational priorities, and to accommodate budget constraints, the survey is representative of host communities living in close proximity to refugees living in refugee villages in KPK, Balochistan, and Punjab. The outer boundary for consideration of the host communities was set by maximum simple Euclidian distance of initially 50km, which was reduced to 30km once the fieldwork began due to extreme remoteness and difficulty accessing of some areas.

Sample representativity

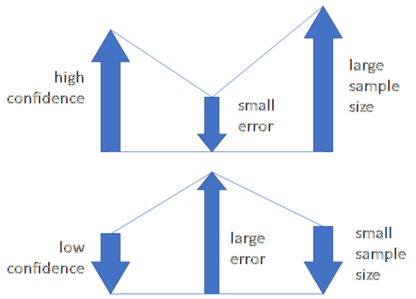
The notion of sample representativity is very often associated directly with the sample size and the notion of statistical power or robustness, reliability of statistical estimates – i.e. how confident are we that the number derived from the data actually represents the reality in the population. However, sample representativity is actually not determined by the sample size but is rather a reflection of the structure of the sample and how similar (or different) it is from the actual population in terms of key socio-economic and geographical strata (groups). Sample size, in turn, determines the precision and statistical power of the estimates from the sample independent from the sample representativity.

Disaggregations and statistical power



The FDS will be representative of the target populations described above, however data disaggregation within each target population is desirable as contexts and realities across the population and programming and support for refugees and hosts varies geographically. The level and number of desired domains over which the data will be disaggregated is an essential sampling parameter as sufficient power and precision must be achieved for each domain. A further essential sampling parameter to define is disaggregations (or analytical domains) and with it the statistical power with which results will be presented for each disaggregation. This disaggregations (or analytical domains) define the sub-groups within a given target population for which analytical results can be derived in a statistically robust way (i.e. with confidence that the results obtained from the data reflect the reality in the population). For example, typical disaggregations of the findings on a country's refugee population as a whole could be by province or by in-camp versus out-of-camp. While greater granularity is of course always desirable, it must be noted that each additional disaggregation necessitates a considerable increase in total sample size – and hence need to be selected cautiously.

A survey's sample size requirement is driven by the required statistical precision to deliver the analytical results in a statistically robust manner –i.e. with high confidence. The higher the confidence in results is required, the larger the sample size needs to be. Sample size calculations are discussed in more detail below. In order to have high confidence in the results produced at a certain disaggregation level, the sample size of that disaggregation needs to meet the required minimum. The size of these analytical groups or disaggregations are referred to in this text as MAES or Minimum Acceptable Effective Size.



This Forced Displacement survey is designed and powered to produce representative estimates of the following sub-groups

- Refugees
    - Khyber Pakhtunkhwa
        - in-refugee village
        - rural at large out-of-refugee village
    - Baluchistan

        - in-refugee village
        - rural at large out-of-refugee village
    - Punjab
        - in-refugee village
        - rural at large out-of-refugee village
    - Metropolitan refugees
        - Islamabad and Karachi combined
- Hosts
    - By province
        - Khyber Pakhtunkhwa
        - Baluchistan
        - Punjab
    - By proximity to refugee villages
        - Living in close proximity to refugee villages

Where more than one first-level disaggregation is listed per target population, these constitute parallel and not cumulative disaggregations.

Table 2. Administrative units in Pakistan

Levels	Punjab, Sindh, Balochistan		KPK		Census	
	Rural	Urban	Rural	Urban	Rural	Urban
Level 1	Province				Province	
Level 2	District				District	
Level 3	Tehsil				Tehsil	
Level 4	Union Council*		Neighbourhood council*		Patwar circle (PC)	Sector
Level 5	NA		NA		Mouza	Block
Level 6	NA		NA		EA	EA

** UC has population of 15,000 + individuals (1,800+ households assuming average size of 8.2)*

Sampling approaches and frames

Refugees

There are three key locations for sampling of refugees:

- rural at-large
- refugee villages,
- metropolitan areas of Islamabad and Karachi.

Given the differences among these three locations, the sampling frames and approaches used to select and reach each sub-population vary.

Context

Decisions of the Government of Pakistan to involuntarily repatriate any undocumented Afghan refugee to Afghanistan, has resulted in heightened protection risk for all Afghani refugees in Pakistan. This heightened protection risk also had important bearing on the final sampling decisions for the Forced Displacement Survey. Joint assessment of the UNHCR Global Survey Team, UNHCR Pakistan protection team, and survey implementation partner (C4ED) led to a more protection sensitive approach to sampling and contacting procedures for the households selected to participate in the survey. Mobile telephone to sampled households was recommended as the initial mode of contact by the survey implementers as any in-field identification of and search for location of households was deemed to be a severe protection risk both for the refugees as well as the interviewer field teams. This decision informed the sampling methodology and defined the approach.

As the possession of the phone is mandatory for the initial contact only, refugees with a valid phone entered into registration system could be selected.

Rural at-large refugees

A large proportion of refugees live in rural locations outside of refugee villages across KPK, Baluchistan, and Punjab. Detailed location information of where the household is residing (i.e. detailed address) is not systematically available in the DRIVE database, although it is expected that particular refugees may be reached with effort as part of the identification activities in the field. Initial contact as mentioned is limited to sampled units with telephone numbers in DRIVE.

DRIVE registration database as the Sampling frame

The sampling frame for the strata of Afghan PoR card-holder refugees in Khyber Pakhtunkhwa, Baluchistan and Punjab was an extract of all refugees (individuals) living in the designated geographic area. The records represent registered individuals who are organised into registration cases as identified by the combination of individual and case IDs. Registration cases approximately resemble a family but are not entirely equivalent to the definition of household as an economic unit used in the household surveys. Each registration case features one individual who is considered the case principal and can be encapsulate one or more families with their heads. The list comprised of the family heads, as identified by UNHCR, was used to construct the final sampling frame.

Admin levels 0 to 3 are standardised and use nationally and internationally compatible classification. Any details at admin level 4 or below are recorded in an open text field stored under admin 6 field within the registration database. Admin 6 field would hold the lowest possible details down to the street address. However, the information in the Admin 6 field is not structured and therefore cannot be used in systematically determining the exact location of the sampled households.

Sampling approach

The Sample for PoR card holders living in at-large in rural areas is drawn systematically randomly stratified by tehsils (admin 3) without any geographical restriction. The drawn sample includes both households with a registered telephone number and those without it. Only households with registered telephone number are considered as eligible for interview attempt.

In order to address remoteness of some areas in Khyber Pakhtunkhwa and Baluchistan and high level of dispersion of refugees in Punjab, tehsils with lowest refugee numbers were under-sampled. Low density tehsils were defined as tehsils with lowest numbers of resident refugees where 10% of the refugee population in each province resided.

These low-density tehsils representing 10% of the population represented 3% of the sample size.

These strata use oversampling as the means to address nonresponse (both due to lack of registered telephone contact number as well as a result of refusal or noncontact outcome of the phone contact attempt) and no replacement protocols are used for this stratum. Assumed response rate is expected at approximately 50% so initial sample drawn is twice as large as the targeted size per stratum. Based on the exhaustion of initial sample and response rates achieved, additional top-up samples were issued. It is important that the full sample has been properly exhausted (all of the contact attempts are used on all the sampled households), and the data collection has not been stopped before that (not stopped upon reaching the required sample size).

- In this stratum the focus was on refugee households with a phone number registered in DRIVE. The phone contacting attempt protocol was set to maximize the response rate. During the successful contact attempt the consent to be interviewed was obtained with detailed description of the location (address) of the household's dwelling.

- If no response is achieved over the phone and for those without the listed phone number, the address listed in the registration data (free-form address field – admin 6) was assessed if the exact address can be identified – with minimal to no involvement of the community facilitators.

- If the address could not be identified, the contact attempt was deemed to be a nonresponse.

Refugees in refugee villages

For the refugee villages, single stage systematic random stratified sampling across all locations is used and using DRIVE registration data as the sampling frame. An area sampling using Google Building footprint data has been considered as a possible approach to sampling for this stratum initially. However, due to the reports of high mobility of the PoR card holders out of the refugee villages and influx of either ACC card holders or those Afghans lacking either of the recognised documents, the population of refugees within the villages is very heterogeneous and an area-based approach would result in high number of ineligible households being selected.

The refugees in the refugee villages are sampled by stratified systematic sampling where the primary stratification criterium was the village and demographic characteristics serve as secondary stratification criteria. Thus, the allocation of the refugees in the sample is proportionate to the recorded population of the PoR cardholders across villages. The initially drawn sample includes both the refugee households with available phone number as well as without.

The RVs offer more security and thus less protection risk for both the refugees (respondents) and interviewers. Hence, alternative contacting strategies could be explored – phone contact and in-filed identification with the help of community leaders. This provides a unique opportunity to better understand potential biases in the out-of-RV sample of refugees, where only telephone contacts were used.

The following protocol was used in the field:

- Sampled households with a valid phone number provided in the sampling frame were initially contacted by phone. Following the same protocol as the rural-large sample of refugees.
- Both the nonrespondents on the initial phone contact as well as refugees with no phone provided (or with an invalid phone contact number) were also attempted to be identified using in-field procedures using key informants from within the RV communities.

Metropolitan refugees

Metropolitan refugees are located in two large cities of Islamabad and Karachi. The main sampling frame for sampling of this stratum is DRIVE registration data. The frame was implicitly stratified by city first, with city sample sizes proportional to their target refugee populations, and systematic random sampling used within each stratum. Further administrative divisions within each city were used to further stratify the sample. Both refugees with valid phone number and those without were drawn, but only those with a valid phone number provided in the frame were contacted and subsequently interviewed. This stratum used oversampling to address the nonresponse.

Hosts

The sample of host communities is divided into three distinct strata:

- Host communities in close proximity to RVs in Khyber Pakhtunkhwa
- Host communities in close proximity to RVs in Balochistan
- Host communities in close proximity to the single RV in Punjab

Sampling of the host population in the three strata was based on the Google Building Footprints database. Google Building Footprints is a database of geocoded building data, that is acquired by processing high resolution aerial photography. The sampling frame included all the building objects from the database located within the 30 km outer boundary of distance from the refugee villages and outside of the boundaries of these RVs. The sample of hosts was selected with systematic random sampling method using probabilities proportional to proximity of the respective buildings to the closest refugee village boundary.

Proximity (distance) estimation was modified (with power transformation) in order to achieve approximate allocation of 50% of the sample within 10 km distance.

Replacement protocols were used for this stratum.

Sample size

Overall approach to sample size calculation and composition is based on the analytical requirements and subsequently identified essential analytical domains and disaggregations as described earlier in this document. This leads to an optimal sample allocation which is based on balanced allocation of sample units per identified analytical domain. The required sample size is therefore estimated at the level of each disaggregation. For this purpose, the Minimum Acceptable Effective Size (MAES) is used. MAES identifies the minimum number of households required for a desired level of precision in each disaggregation.

The proposed sample size per MAES will be calculated using the following formula:

$$n = \frac{z^2 \times p(1-p)}{\alpha^2}$$

where

n = net sample size
z = z-score
p = proportion of population with a given trait
α = margin of error

Effective sample size per MAES depends on the sampling method proposed for a particular stratum. Proposed sampling method for most of the explicit strata for the FDS in Pakistan will use single stage sampling and thus would require smaller size per MAES of $n = 500$.

The table below summarizes a suggested allocation as expressed in MAES.

Table 3. Sample allocations per MAES (explicit stratum)

	MAES	Clustered	Frame & Approach	listing	n
Refugees					
In RVs in KPK	1	No	DRIVE	No	500
Outside of RVs in KPK	1	No	DRIVE	No	500
In RVs in Balochistan	1	No	DRIVE	No	500
Outside of RVs in Balochistan	1	No	DRIVE	No	500
In RVs in Punjab	1	No	DRIVE	No	500
Outside of RVs in Punjab	1	No	DRIVE	No	500
Metropolitan in Islamabad and Karachi	1	No	DRIVE	No	500
Hosts					
KPK close to RVs	1	No	Building footprints	No	500
Balochistan close to RVs	1	No	Building footprints	No	500
Punjab close to RVs	1	No	Building footprints	No	500
Total	10				5,000

Replacement protocol

In order to address nonresponse in the refugee samples in RVs and among the hosts a replacement protocol has been put in place. Replacements as the protocol to address potential unit nonresponse are often discouraged by producers of official and academic as they can result in biased samples if used inappropriately. However, replacement protocols are needed in surveys where certain sample sizes need to be achieved in order to ensure target statistical power and the costs of the data collection need be optimised. Use of replacement protocols enables cessation of field activities once the target samples size is reached without compromising the selection probabilities. This enable the FDS field teams to ensure that the target sample sizes are achieved while managing costs of data collection. The initially drawn sample is exactly the size of the target sample, and it will always be fully exhausted during data collection.

Replacement samples were drawn together with the main sample and selection into a replacement sample was randomised. Issuing of replacement was further randomised at the level of the explicit sampling stratum and is strictly controlled – i.e. issuing of replacement households is managed by the central survey coordination team of the implementing agency. Further, the use of households from the replacement pool is strictly documented.

As the replacement samples were drawn together with the main sample the two samples share the same selection probabilities.

The replacement sample is included together with the main sample and encoded into the Kobo Questionnaire form. Any household belonging to the replacement sample are protected with a validation code. If an interviewer selects a replacement sample he is prompted for a validation code, which is matched to the encoded value. Only if the interviewer is provided with the validation code, the questionnaire can advance to the interview. This setup supports flexibility as both the main and replacement samples are readily available at interview time, while maintaining strictly controlled issuing of the replacements – issuing of validation codes is under the control of the survey coordination team.

Sample adjustment

As part of the data processing tasks, the sample was adjusted using weighting procedures. All the weights used in FDS data are analytical weights – i.e. total sum of weights is equal to the sample size. The weights used for analysis are composite weights comprising of basic sampling weights as well as structural adjustment weights. Sampling weights correct for unequal probabilities of selection across different strata, while structural adjustment weights adjust to basic population structures such as geographic distribution. Structural weights are also used to adjust the sample of balanced size sample strata to population proportions in order to derive national estimates.

FDS data analysis does not use population weights – i.e. weights that sum up to the population totals and thus enable the analyst to estimate true population numbers for indicators and across population groups. FDS is not meant to be the source of population data and as such does not provide weight for such estimations.

Process of weight estimation

As mentioned, final analytical weights in FDS are composite weights composed of base weights and structural adjustment weights. The weight estimation closely follows the sampling methods used in the selection process as well as adjustments of protocols, if any, as implemented in the field. In the strata where single stage systematic random selection is used there are four basic steps that are followed:

1. In the first step the base probabilities of selection are estimated separately for each explicit stratum (sample allocation stratum - Table 2). Selection probability is the calculated as follows:

   $$p_i = \frac{n_s}{N_s}$$

   Where n_s is the total number of sample units drawn into both the main and replacement samples and N_s is the total number of units in the frame. The base weight is the inverse of the selection probability

   $$w_i = \frac{1}{p_i}$$

2. In the second step the initial base weights are rescaled to the realised sample size of respondents (r_s).

   $$w_i^* = \frac{w_i}{\sum w_i} \times r_s$$

3. In the third step structural adjustments are made. As the full information matrices on geographical distributions of the sampled population exist, poststratification is used to adjust the weights to the correct population proportions.

Due to in-field adjustments to issuing of replacements, the weight estimation process assumes that the sampling has been carried out at sub-stratum level. The weight estimation reflects that. The sub-strata units remain represented proportionially.

Google footprint samples of hosts based on proximity

The samples of host in proximity to the camp in the Extreme north and settlements in the Est, are approximately self-weighting according to the selection gravity coefficient. Selection gravity coefficient is power adjusted (6th root) simple Euclidean distance from the building to the closest settlement (camp) boundary. As the selection was carried out over two explicit strata (0 – under 10 km & 10km to 20km), there may be a need to structurally adjust the sample according to unrestricted (not split over two strata) gravity coefficient. Self-weighted sample represents the population that leaves in proximity of the refugees (selection gravity coefficient based on proximity favours those living closer in terms of selection).

$$dist_{ijk} = |b_{ij} - P_{ik}|$$

$$dist^{*}_{jk} = \sqrt[6]{dist_{jk}}$$

$$prox^{*}_{ijk} = round1,000 \times \frac{\max_{1 \leq j_k \leq J_k}(dist^{*}_{ijk} - dist^{*}_{ijk})}{\max_{1 \leq j_k \leq J_k} dist^{*}_{ijk}}$$

$$p = \frac{1}{prox_{ijk}}$$

Number of buildings adjustment for building footprint samples

Weights for building footprint samples of hosts and refugees in settlements in the Est are further adjusted for the number of buildings the household owns (lives in). Adjustment coefficient is one (1) over the number of buildings owned. The building ownership is self-reported as part of the interview process. Households who own multiple building have higher probability of selection, which needs to be adjusted as part of the unequal probability adjustment weight estimation.

Additional weights are estimated for the analysis of individual datasets (random adult, random child under 5 and random woman who gave birth in the last 2 years). Individual weights are composite weights of household selection weights with within household selection weights.

Nonresponse adjustment

Due to adverse conditions on the ground and challenges in locating the refugee populations, the response rates vary greatly across the sampling strata and thus nonresponse adjustment component of the composite analytical weights was deemed necessary.

Based on the available information on the final dispositions of the sampled units and structure of the target population based on registration data (DRIVE), it was decided that the best method for nonresponse adjustment was estimation of nonresponse propensities using probit models. The models used the socio-demographic characteristics of respondents and nonrespondents available in drive to predict response, this included not only demographic features of the case, like household size, family composition, age of the head, education, etc, but also a selection of economic characteristics such as employment.

Structural adjustment

For structural adjustment, the source of the target population distribution is the registration database DRIVE. It needs to be noted that the population distribution based on the registration data is not deemed to be the most reliable, particularly based on the observed differential survey outcomes across different strata. This is predominantly due to high mobility of the population as well as large-scale waves of returns to the host country (Afghanistan). Sampled individuals were often not found in locations where they were registered. During the survey data collection, it could not be established whether the nonrespondents have merely moved internally or have left the country entirely. Despite these concerns DRIVE database remains the most reliable existing source of population vital statistics for refugees and asylum seekers. As no other source of population data is available, DRIVE remains the source for the estimation of the nonresponse calibration weights.

Use of weights in analysis

Any analysis using FDS data should use the supplied or equivalently estimated weights in order to derive unbiased statistics.

The use of weights depends on the target population of a given indicator as well as on the purpose of analysis. The following analytical purposes are considered:

- estimation of national or sub-national indicators or models – structurally adjusted national weights (proportion of sampling strata is adjusted according to national population structure)
- comparative analysis of sampling strata – structurally adjusted strata weights (size of the strata is not adjusted to population proportions and left at the size sampled to maximise statistical power of analysis).

The following target populations are defined:

- Population of refugee households – household weight
- Population of all household members in the households – household weight
- Population of refugees aged 15 and above – individual weight
- Population of children under the age of 5 years – child weight
- Population of women who gave birth in the last two years – woman weight

Variance estimation and analysis

Variance estimation will be facilitated by the use of survey design specification commands imbedded in the Statistical software like Stata or R. In Stata `svyset` command will be used to specify the survey design, associated weights and strata. For analysis `svy`: prefix will be used to estimated complex variances. Similar functions exist in R.