World Bank R2 - Sampling Note

Sampling Frames

This survey attempted to reach 11,738 households by mobile phone. The Data on Demand team successfully completed 5,005 forms, a response rate of ~46% of the households for whom valid numbers existed. The phone numbers used in this survey were assembled from three prior IDinsight projects, and from an impact evaluation of the National Rural Livelihoods project conducted by the Ministry of Rural Development Each of these surveys sought to represent distinct populations, and employed idiosyncratic sample designs and weighting schemes. Key features of each dataset are summarised below:

- **CIFF Poshan Abhiyan Monitoring (CIFF)**
    - **Description:** Two-stage stratified cluster random sample. Seven districts in Rajasthan and five districts in Jharkhand were randomly chosen from strata designed to capture relevant heterogeneity within each state. Within each district, 35 polling stations were selected with probability proportional to size (PPS) from assembly constituency (AC) strata, and then 15 households were chosen from each polling station.
    - **Round 2 sample size:**
        - Rajasthan: 1,500 attempted, 930 complete
        - Jharkhand: 1,500 attempted, 890 complete
    - **Representativeness:** full rural populations of Rajasthan &Jharkhand
    - **Time frame:** drawn in winter 2019-2020
    - **Sample frame:** voter rolls
    - **Stratified:** yes, at the AC level
    - **Clustering:** at the polling station (primary sampling unit) level
    - **Weights:** probability weights defined for original sample

- **BMGF Poshan Abhiyan SBCC Monitoring (SBCC)**
    - **Description:** Two-stage stratified cluster random sample. In Andhra Pradesh 200 villages/wards were chosen with probability proportional to size from strata defined by a range of socio-economic indicators. Six households were then chosen from the registry of a FLW whose catchment covered the selected village/ward.
    - **Round 2 sample size**:
        - Andhra Pradesh: 549 attempted, 281 complete

    - **Representativeness:** representative at the state level of all households with pregnant or lactating mother listed by an ASHA/AWW
    - **Time frame:** drawn in fall 2018
    - **Sample frame:** Frontline workers' (ASHA & AWW) registries
    - **Clustered:** at the village/ward (primary sample unit) level
    - **Weights:** probability weights defined for the original sample

- **State of Aadhaar Report (SOAR)**
    - **Description:** Three-stage stratified cluster random sample. Three districts in Andhra Pradesh were randomly chosen with PPS. Within each district, 20 polling stations were selected with probability proportional to size (PPS) from assembly constituency (AC) strata , and then 10 households were chosen from each polling station.
    - **Round 2 sample size**
        - Andhra Pradesh: 456 attempted, 230 complete
    - **Representativeness:** Representative of AP at the state level
    - **Sample frame:** Voter rolls
    - **Clustered:** at the polling station (primary sampling unit) level
    - **Weights:** probability weights defined for the original sample

- **National Rural Livelihoods Programme (NRLP)**
    - **Description:** Sample was collected by the World Bank and covers 9 states that were part of the National Rural Livelihoods Programme. Number of households sampled were 2398 in UP, 4524 in Bihar and 2877 in Madhya Pradesh.
    - **Sample size:**
        - Uttar Pradesh: 1,899 attempted, 778 complete
        - Bihar: 2,658 attempted, 1,073 complete
        - Madhya Pradesh: 2,339 attempted, 823 complete
    - **Representativeness:** Representative of SHG membership in states. Rural districts selected from strata to reflect a range of outcomes.
    - **Sample frame:** Undefined
    - **Clustered:** at the village (primary sampling unit) level
    - **Weights:** not defined in the original survey

Challenges

The above features of the sample impose limits on the inferences that we can draw and imply analytic challenges.

1. *Representativeness*: For each state, the sample is not necessarily formally representative of the full state population. Of the six states, only **Rajasthan and Jharkhand** are represented by an (arguably) unbiased sample for the full rural population. **Andhra Pradesh** is covered partially by a representative sample (SOAR), but that sample is pooled with an non-representative one (SBCC). **Bihar, Madhya Pradesh, and Uttar Pradesh** are covered by the NRLP sample, which captures data from rural districts where the program was implemented. These districts may have had less favorable outcomes *ex ante*, but in some instances outcomes may be improved due to the prevalence of SHGs.

   The representativeness problem is compounded when we pool the data for cross-state analysis. Three main problems arise. First, the pooled sample is not representative of any population in particular, but rather it inherits the biases of each state sample, and represents some generalized, amorphous rural population. Second, most states in the sample have roughly equal final sample sizes, but the states vary widely in total population. Each state's observations should be reweighted to reflect the imbalance. Finally, absence of a unified sampling strategy, and therefore incompatible probability/sample weight, make it difficult to analyze the data as a whole.

2. *Noncoverage*: The sample frame comprises households with mobile phones. Phone owners may differ from non-phone owners in ways (socioeconomic status, SC/ST status, remoteness of a village) that are correlated with outcomes of interest.
3. *Nonresponse*: Similarly, nonrespondents may differ systematically from respondents.To the extent that nonresponse is correlated with outcomes of interest, estimates may be biased. Nonresponse also decreases precision by reducing the sample size.

Base Weights

Base weights are merged into the cleaned dataset. Base weights reflect a probability of selection into the original sample, and can be interpreted as an expansion factor to some population. State-wise details follow:

1. ***Bihar, Madhya Pradesh, Uttar Pradesh*****:** As noted above, probability weights are not included in the NRLP sample. We assign even base weight.
2. ***Andhra Pradesh*****:** The source datasets provide base weights, which expand to the state population (SOAR) or the population of households listed on ASHA/AWW rosters (SBCC).
3. ***Jharkhand, Rajasthan*****:** The source dataset provides base weights, which expand to full district populations. These districts were chosen with PPS from geographic strata to represent the rural population of the two states.