Innovative sampling strategies for difficult to survey population

Haoyi Chen, Coordinator
Inter-Secretariat Working Group on Household Surveys
Difficult-to-survey population

- **Difficult to identify or to reach populations**: rare and relative rare population groups; marginalized, mobile and elusive, hidden sub-groups of the general population such as homeless population, undocumented migrants etc.

- **Difficult to survey populations**: vulnerable populations requiring explicit consent from or to be substituted by a caretaker, parent, or guardian, e.g. young children, prisoners, institutionalized persons, and unself-sufficient elders.

- **Difficult to count or to measure survey variable(s)**: stigmatized and sensitive study variable. These might include sexual orientation, unreported work, use of drugs and mortality, ethical and privacy concerns.
Sampling to leave no one behind – Inter-Secretariat Working Group on Household Surveys

- Introduction
- Chapter 1. Policy relevance for SDG data disaggregation and implication on sampling
- Chapter 2. Population groups covered by the Guidance
- Chapter 3. An overview of sampling strategies
- Chapter 4. An overview of challenges and strategies in sampling certain population groups
- Chapter 5. Guidance on sampling various population groups: concepts and definitions, sampling methods and country practices
  - Income, poor and extreme poor
  - Sex (gender)
  - Age (including children and older persons)
  - Race and ethnicity (including people with African descendants)
  - Migratory status (including migrants, forced displaced)
  - Disability
  - Geographic location (including urban/rural)
  - LGBT (gender identity and/or sexual orientation)
  - Drug users
  - Homeless
  - Prisoners
Sample size for adult mortality estimates in surveys

- **RHD:**
  - Level of mortality
  - Precision required for the estimates
  - Length of reference period

- **SSH – additional dimensions**
  - Number of siblings being listed
  - Probability that these siblings 15+ during the reference period

What about further disaggregation, urban/rural, age, sex, education?
Sample size for adult mortality estimates in surveys (2)

- Rwanda, 2010: <0.05 deaths identified per interview (SSH)
- Bendavid et al. (2011): with national representative survey data from eight sub-Saharan countries
  - More than 300,000 households
  - RHD
  - 1,233 deaths above age 60 (just under 20% of all deaths), or about 75 deaths per country, disaggregated by sex

- Most surveys are not designed to collect adult mortality:
  - Larger confidence intervals compared to other indicators in the survey
  - Longer reference period, typically 6-8 years prior to the survey (SSH)
Innovative sampling strategies - oversampling

MICS (Megill D., Khan SM., and Hancioglu A. (2018))

- Oversampling of households with under-fives follows a similar approach to the standard MICS approach, wherein sample sizes and the number of EAs are first calculated, although it is necessary to consider the oversampling strategy on the number of under-fives expected in the sample. Sampling specialists first examine the number of under-fives that a usual sample would yield and then examine various oversampling rates to determine the gains of the strategy.

- After the census enumeration areas are selected (i.e. the first stage sampling), for the listing of the households, the listing form is designed to identify which households have children under-five. Then, the surveys select these households with a higher sampling rate compared to the households without children under-five.

Innovative sampling strategies – oversampling (2)

- Challenges with oversampling:
  - Additional listing -> more costly!
  - False positive/negative
  - Additional weighting needed -> higher design effect -> less precision

- In some instances, oversampling can be made more efficient if the rare event is clustered such as migrants or ethnic minorities
  - Is it the case for adult mortality?

- Not tested in the adult mortality context
Innovative sampling strategies – network sampling

**How Many People Do You Know?: Efficiently Estimating Personal Network Size**

Tyler H. McCormick, Matthew J. Salganik, and Tian Zheng

In this article we develop a method to estimate both individual social network size (i.e., degree) and the distribution of network sizes in a population by asking respondents how many people they know in specific subpopulations (e.g., people named Michael). Building on the work of Moskowitz et al. (2001) and Saha et al. (2001), we introduce the scale-up method which corrects for known problems with previous approaches. As expected, our method also provides estimates of the size of social networks across population subgroups. We assess the performance of our method using data from the Pew Research Center’s “How Many People Do You Know?” study. Results show that our method has lower variance than previous approaches and is more accurate in estimating the distribution of network sizes. The estimates from the sample scale-up model may not represent the population in the same way as the estimates from our more complex model for estimating network sizes.

**KEY WORDS**

- Latent network size estimation
- Network survey design

**SCALE-UP METHODS AS APPLIED TO ESTIMATES OF HEROIN USE**

Charles Kadziu, Peter D. Kellermann, H. Russell Bernard, Andrew A. Beveridge

The feasibility of using the network scale-up method to estimate heroin use is described. An example sample was asked “How many people do you personally know who use heroin, and how many in other subpopulations?” Using this approach, we find that the average number of persons who use heroin is almost 3 times higher than the rate of such self-report. Estimates of the subpopulation are compared with known subpopulation sizes to assess the feasibility of the model. Data from the 1999 survey evaluating the “Fighting Back” substance prevention program. Fortwien ideas with clear political boundaries were used (n=5660). Heroin use varied from city to city. Rates estimated for heroin use (correlated with R=0.7) are less for respondents from cities. The average ratio between the known populations and the estimates is 0.84. Members of each subpopulation, especially drug users, tended to know more people within their own subpopulation.

**ABSTRACT**

Estimating size of hidden or hard-to-count populations is an important problem in public health. For example, estimates of the sizes of populations at highest risk for HIV and AIDS are needed for designing, evaluating, and allocating funds for treatment and prevention programs. A promising approach to size estimation, relatively new to public health, is the network scale-up method (NSU), involving two steps: estimating the personal network size of the members of a random sample of a total population, and this, with the information, estimating the number of members of a total population. The method, known as the scale-up method, has been used to estimate the size and distribution of network sizes. We describe the method, including two approaches to estimating personal network size (survey and sample). We discuss the strengths and weaknesses of each approach and provide examples of international applications of the NSU, in public health. We conclude that samples of a population as well as a method to uniquely identify which individuals were recruited in more than one sample. Synthetic estimates and suitable estimators for the size of hidden or hard-to-count populations of the network scale-up method (NSU). We describe the background of the method, its results in public health, and an evaluation of its strengths and weaknesses. Finally, we report a literature review method for improving the method’s utility for programming and planning, based on the consensus of an expert panel ([see online supplementary appendix 1]).

**Practice of Epidemiology**

Assessing Network Scale-up Estimates for Groups Most at Risk of HIV/AIDS: Evidence From a Multiple-Method Study of Heavy Drug Users in Caripito, Brazil

Matthew J. Salganik, Dinah G. Faddis, Nellie Bartone, Alexander H. Bello, Laura B. Miller, and Francesca J. Destro

Network sampling – Nigeria drug use survey

Self reported
- Cannabis use 1.4%
- Heroin use – 0.03%
- Non medical use of opioids (tramadol) 3.8%

National estimate
- Cannabis use – 10.4%
- Heroin use – 0.1%
- Non medical use of opioids (tramadol) 4.8%
Network sampling – adult mortality

Network sampling – adult mortality (2)

**Rwanda 2010-11**

- 13,761 women who were asked to report on their siblings
- Sibling estimates of death rate: 7-year reference period before interview
- Network: 1-year

![Deaths per interview](chart.png)
Challenges with network sampling

- Long questionnaire
- More difficult to obtain SES
- Network might change during crisis?
Increase effective sample size for adult mortality measurement

- Adult mortality is a rare event and difficult to capture (dead!)
- Statistically, this requires larger sample size
- Innovative sampling strategies can help but also have their limitations; and some requires more testing in the adult mortality context
- How does this fit into your household survey system?
  - Adding modules to continues household surveys?
  - Integrating different sources: Survey with other data sources (census, survey, geospatial, CRVS, burial)?
Questions?

chen9@un.org