Estimating The Size Of Hidden Populations

CNSTAT Workshop on Improving Health for Small Populations, January 18-19, 2018

Katherine R. McLaughlin
Assistant Professor, Department of Statistics

mclaugka@oregonstate.edu

http://stat.oregonstate.edu/~mclaugka/

Oregon State University

Outline

- What is a hidden population?
- Challenges of estimating the size of hidden populations
- Population size estimation methods
  - Capture-recapture methods
  - Object and service multipliers
  - Network scale-up methods
  - Successive Sampling-Population Size Estimation (SS-PSE) for respondent-driven sampling

Hidden Populations

- Hidden populations may also be called:
  - Hard-to-reach
  - Hard-to-sample

- Members of hidden populations may engage in behaviors that are sometimes illegal or stigmatized and thus may tend to avoid disclosure of their membership and be unwilling to participate in surveys.

- Examples:
  - Key populations at high risk for HIV, such as female sex workers (FSW), people who inject drugs (PWID), men who have sex with men (MSM)

Why Populations Size Estimation?

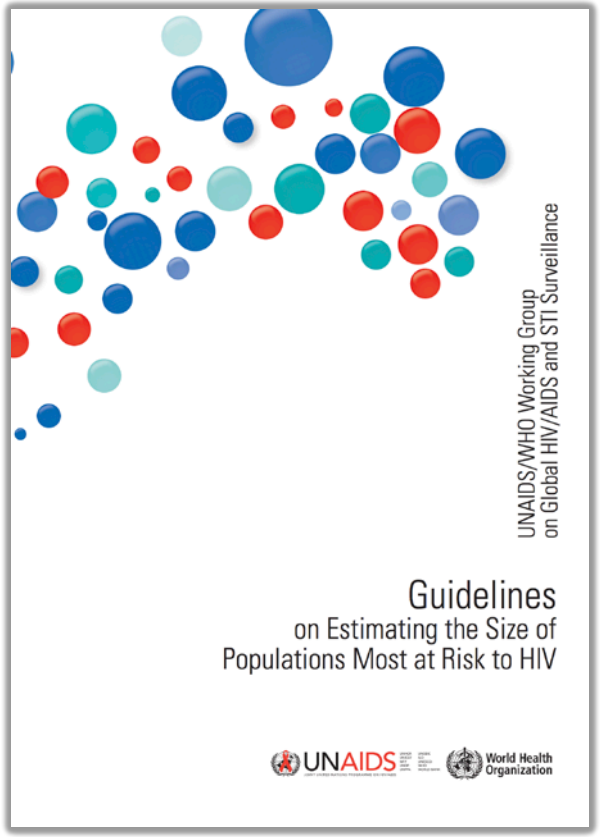
- To assess the existence or magnitude of an issue relating to the population
- To assess how resources should be allocated for better program planning and management
- To aid other estimation methods for these populations, which may require knowledge of *N*
- If repeated over time, to assess population dynamics

Challenges of Population Size Estimation

- A sampling frame may not exist
- Members of hidden populations may not identify themselves and members of the general population may not know whether or not their friends are members of the hidden population
- Those who participate in a sample may be different than those that do not
  - May be more likely to observe people who are more visible/ highly connected
  - Non-participants may be more isolated or even completely separate from those who do participate
- Populations are dynamic – both in time/space, and membership
  - Timing matters for methods that rely on two samples

Current Methods for PSE

- No “gold standard” currently exists and many methods have been used, each with different strengths and weaknesses
- The particular approach chosen will depend on the population of interest and resources available



UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance (2010). Guidelines on Estimating the Size of Populations Most At Risk to HIV.

How to Reach a Hidden Population

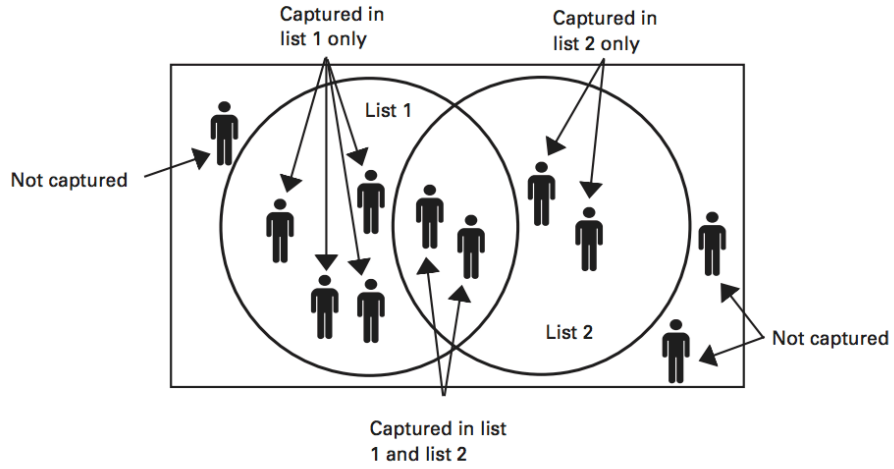
If we have no sampling frame, what do we do?

- Rely on a general population survey.
  - Inefficient if we only are interested in the hidden population.
  - Individuals may be unlikely to disclose status as a member of hidden population, and people may not know their friends' statuses.
- Venue-based / time-location sampling: identify locations where members of a hidden population are likely to congregate. Sample locations instead of people.
  - May be difficult to identify a list of venues.
  - May miss individuals who do attend venues.
- Respondent-driven sampling: identify a few "seed" members of the hidden population, use restricted peer-recruitment to grow sample chains.
  - May be biased by initial choice of seeds, volunteerism, dependence between individuals

Capture-Recapture

- Procedure
  1. Map all the sites where the population can be found
  2. Go to the sites and “tag” all the members of the population at the site
  3. Return to the sites at a later date and retag all members of the population
  4. Record size of each sample and overlap

Figure 3.2. Illustrating the capture-recapture method



Guidelines on Estimating the Size of Populations Most At Risk to HIV, p.17.

$$\widehat{N} = \frac{n_1 n_2}{m}$$

- Simplistic and requires many assumptions, e.g.
  - Every member has an equal chance of being sampled
  - Matching is reliable
  - The two samples are independent

Multiplier Methods

- Service Multiplier Method (SMM) or Unique Object Multiplier Method (UOM)
- Relies on two sources of data
  1. A count of population members who received some service (e.g. attended a clinic or program) or object (unique, memorable)
  2. A representative survey of the population, such as RDS. In the survey, ask each individual if they received the service or object

$$\widehat{N} = \frac{\#\ of\ people\ receiving\ the\ service\ or\ object}{\%\ of\ sample\ who\ reported\ receiving\ the\ service\ or\ object}$$

Multiplier Methods

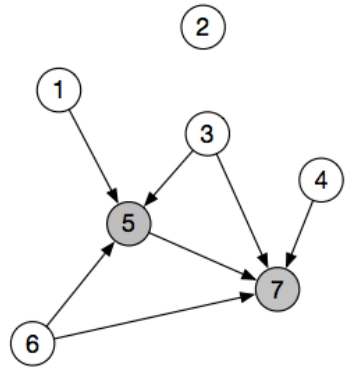
- Challenges
  - Requires that the two data sources be independent
  - Obtaining a random sample of a population lacking a sampling frame
  - Timing between the service/object distribution and the sample
  - Everyone receiving the service/object must be a member of the hidden population

Network Scale-Up Method (NSUM)

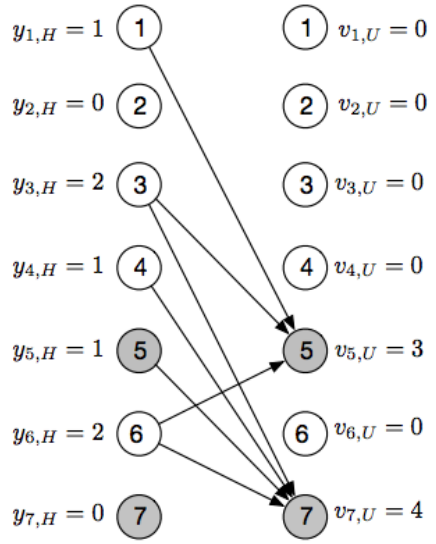
- Procedure:
  1. In a general population survey, ask how many individuals each person knows and how many of those are in the hidden population.
  2. Proportion of respondents' contacts who are members of the hidden population is assumed to be equal to the population proportion. Multiply this by the known general population size.
- Requires the assumption that people in the general population are aware of whether or not their network members are members of the hidden population.
- Assumes network connections are formed at random.

Network Scape-Up Method (NSUM)

- Generalized scale-up estimator
  - Relies on both a general population survey and a hidden population survey (RDS)
  - Total out-reports equals total in-reports
  - Still assumes that hidden population members have aggregate awareness about visibility



(a)



(b)

Feehan, D. M. and Salganik, M. J. (2016). Generalizing the Network Scale-Up Method: A New Estimator for the Size of Hidden Populations.

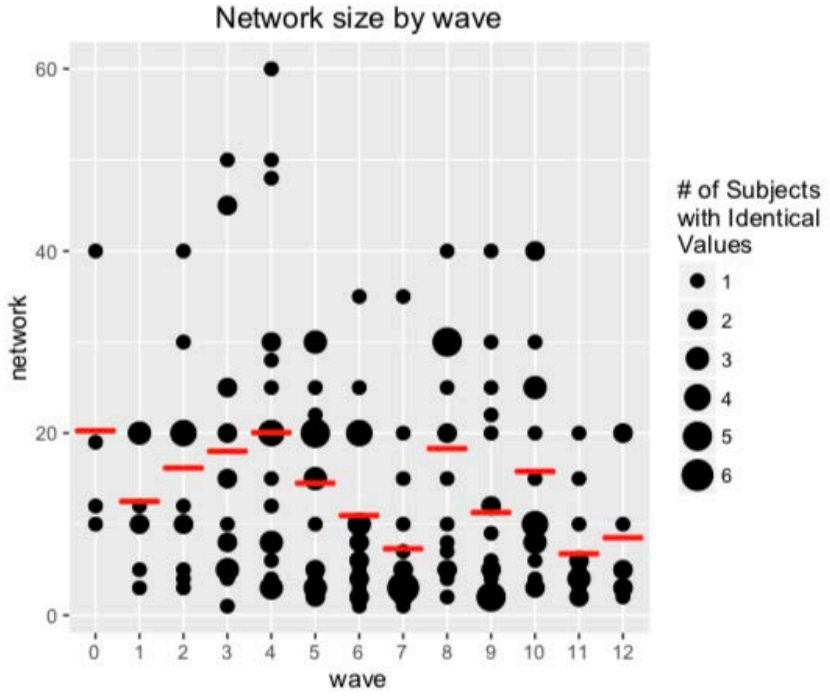
Successive Sampling-Population Size Estimation (SS-PSE)

- Use with data collected via one RDS study
  - Cost-effective
  - Can be done retroactively
- Bayesian framework where prior information about population size can be incorporated
  - Good way to combine data from different sources
  - Statistical model for uncertainty in estimates

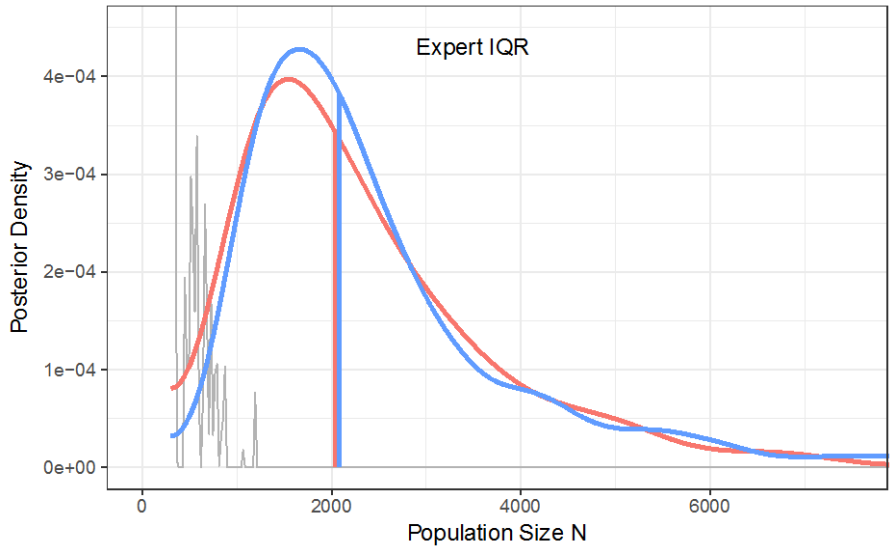
Handcock, M. S., Gile, K. J., and Mar, C. M. (2014, 2015)

Successive Sampling-Population Size Estimation (SS-PSE)

- Conceptual overview
  - People that are more *visible* (tend to have larger network size) are more likely to be sampled in RDS, and be sampled earlier
  - Consider network size by wave
    - If the frequency of larger network sizes decreases over RDS waves, the population is likely being depleted
      - Population size likely larger
    - If the frequency of larger network sizes does not decrease, there are still many people who have not yet been sampled
      - Population size likely smaller



Successive Sampling-Population Size Estimation (SS-PSE)



Example posterior distribution for population size

- Challenges:
  - Network sizes in RDS may not contain a lot of information about population size
  - Relies on quality of RDS data
  - Possibility of inconsistent expert prior beliefs

Future Directions for PSE Work

- Further sensitivity analysis, validation, and diagnostics for existing methods
- Further work on uncertainty estimates for existing methods
- Methods that incorporate multiple estimates
- Opportunities to develop new methods that incorporate
  - New technology
  - Social media data

References

- Fearon, E., Chabata, S. T., Thompson, J. A., Cowan, F. M., and Hargreaves, J. R. (2017). Sample Size Calculations for Population Size Estimation Studies Using Multiplier Methods With Respondent-Driven Sampling Surveys. *JMIR Public Health and Surveillance*, 3(3):e59.
- Feehan, D. M. and Salganik, M. J. (2016). Generalizing the Network Scale-Up Method: A New Estimator for the Size of Hidden Populations*. Sociological Methodology*, 46:153–186.
- Handcock, M. S., Gile, K. J., and Mar, C. M. (2014). Estimating Hidden Population Size Using Respondent-Driven Sampling Data. *Electronic Journal of Statistics*, 8(1):1491–1521.
- Handcock, M. S., Gile, K. J., and Mar, C. M. (2015). Estimating the Size of Populations at High Risk for HIV Using Respondent-Driven Sampling Data. *Biometrics*, 71(1):258–266.
- Johnston, L. G. (2013). Introduction to Respondent-Driven Sampling. World Health Organization, Geneva, Switzerland. Available at http://applications.emro.who.int/dsaf/ EMRPUB_2013_EN_1539.pdf.
- Johnston, L. G., McLaughlin, K. R., El Rhilani, H., Latifi, A., Toufik, A., Bennani, A., Alami, K., Elomari, B., and Handcock, M. S. (2015). Estimating the Size of Hidden Populations Using Respondent-driven Sampling Data: Case Examples from Morocco. *Epidemiology*, 26:846–852.
- Johnston, L. G., Prybylski, D., Raymond, H. F., Mirzazadeh, A., and McFarland, W. (2013). Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations: case studies from around the world. *Sexually Transmitted Diseases*, 40(4):304–310.
- Maltiel, R. A., Raftery, A. E., McCormick, T. H., and Baraff, A. J. (2015). Estimating Population Size Using the Network Scale Up Methods. *The Annals of Applied Statistics*, 9(3):1247–1277.
- UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance (2010). Guidelines on Estimating the Size of Populations Most At Risk to HIV. World Health Organization, Geneva, Switzerland. Available at http://apps.who.int/iris/bitstream/10665/ 44347/1/9789241599580_eng.pdf.