Network scale-up method to estimate the size of hard-to-count groups

Matthew J. Salganik
Department of Sociology
and Office of Population Research
Princeton University

August 24, 2009

This research is supported by UNAIDS and the Brazilian Ministry of Health.

Introduction

There are an estimated 33 million people worldwide living with HIV/AIDS. In most countries, the disease is concentrated in three high risk groups:

- injection drug users (IDUs)
- commercial sex workers (CSWs)
- men who have sex with men (MSMs)

These populations are often called hidden or hard-to-reach.

Better information about the size, behavior, and disease prevalence in these populations can be used to understand and control the spread of HIV/AIDS.

Previous population size estimation methods

Previous methods for estimating the sizes of hidden populations most at-risk for HIV/AIDS:

- Census and enumeration methods
- Population survey methods
- Multiplier methods
- Capture-recapture methods

For an excellent review see UNAIDS (2003).

Introduction

Network scale-up method (Bernard, Killworth, McCarty, et al.) is designed to answer questions like:

- How many sex workers are there in Moscow?
- How many men who have sex with men are there in Mexico?
- How many heavy drug users are there in Curitiba?

Introduction

In this talk I will . . .

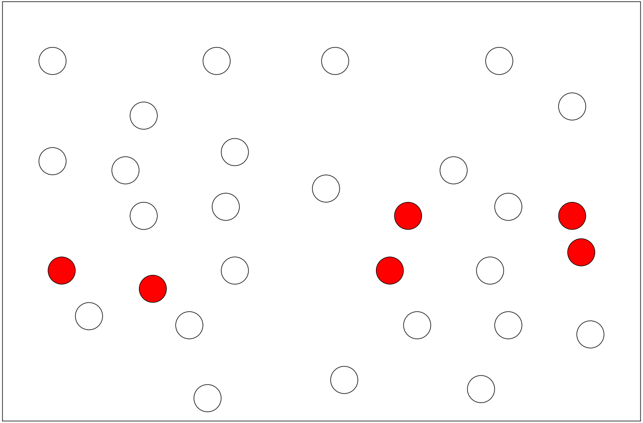
1. Explain how the network scale-up method works
2. Discuss two common approaches for estimating social network size
3. Review strengths and weaknesses
4. Describe the study that is happening in Curitiba
5. List papers that provide additional information

Intuition

Intuition behind network scale up estimate is that peoples' social networks are, on average, representative of the population.

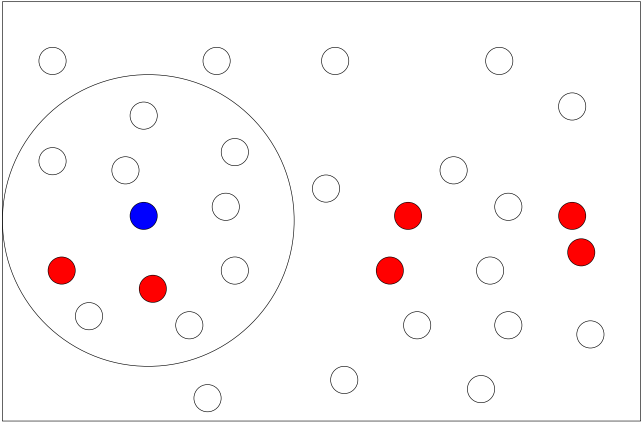
Intuition

Intuition behind network scale up estimate is that peoples' social networks are, on average, representative of the population.



Intuition

Intuition behind network scale up estimate is that peoples' social networks are, on average, representative of the population.



Network scale-up method

Estimated number of IDUs $= \frac{2}{10} \times 30 = 6$

This estimate requires three pieces of information:

1. number of IDUs known (collected in survey): 2
2. network size of respondent (estimated from survey): 10
3. number of people in the entire population (known): 30

Network scale-up method

With multiple survey respondents the estimator is (Killworth et al., 1998)

$$\hat{e}_j = \left(\frac{\sum_{i=1}^{n} m_{ij}}{\sum_{i=1}^{n} c_i} \right) \times t \tag{1}$$

where:

- $\hat{e}_j$ is estimated size of hidden population j
- m_{ij} is the number of people in population j known by person i
- c_i is the the number of people known by person i
- t is the size of the entire population

Network scale-up method

Note that:

$$\hat{e}_j = \left(\frac{\sum_{i=1}^{n} m_{ij}}{\sum_{i=1}^{n} c_i} \right) \times t \neq \left(\sum_{i=1}^{n} \frac{m_{ij}}{c_i} \right) \times t \tag{2}$$

Since we are only estimating sums, this should make the estimates more robust.

How does the network scale-up method work in practice?

- Requires a random sample of the general population; does not require contact with hidden population
- Respondents are asked:
  - How many people do you know who are drug injectors?
  - How many people do you know who are sex workers?
  - How many people do you know who are men who have sex with men?
  - A set of questions to estimate respondent's social network size
- Takes about 5 to 10 minutes per respondent (McCarty et al., 2001)

Know problems: Estimating personal network size

Estimating personal network size (number of people known) is difficult. Median is thought to be between 300 and 750 so complete enumeration is impossible. Two known methods (McCarty et al., 2001):

- Back-estimation method
- Summation method

Estimating personal network size: Back-estimation method

Ask respondents how many people they know in subpopulations of known size (Killworth et al, 1998):

- How many people do you know named Michael?
- How many people do you know who own a Mercedes-Benz?
- How many people do you know who are medical doctors?

Estimating personal network size: Back-estimation method

Ask respondents how many people they know in subpopulations of known size (Killworth et al, 1998):

- How many people do you know named Michael? 5

If there are 5 million Michaels in U.S., we estimate that network size $= \frac{5}{5 \text{ million}} \times 300 \text{ million} = 300$.

Estimating personal network size: Back-estimation method

Ask respondents how many people they know in subpopulations of known size (Killworth et al, 1998):

- How many people do you know named Michael? 5

If there are 5 million Michaels in U.S., we estimate that network size $= \frac{5}{5 \text{ million}} \times 300 \text{ million} = 300$.

- **Strength**: Fits in a statistical framework allowing for quantification of uncertainty; allows for "reality check" of estimates of hidden population size
- **Weakness**: Accuracy of responses is unknown; possible biases introduced by the set of questions; requires relatively accurate administrative records

Estimating personal network size: Summation method

Ask respondents how many people they know in a set of categories and then sum results (McCarty et al, 2001). For example,

- Immediate family, other birth family, family of spouse or significant other, coworkers, best friends, friends through hobbies, neighbors, etc.

Estimating personal network size: Summation method

Ask respondents how many people they know in a set of categories and then sum results (McCarty et al, 2001). For example,

- Immediate family, other birth family, family of spouse or significant other, coworkers, best friends, friends through hobbies, neighbors, etc.

- **Strength**: Doesn't require any administrative records; simple to understand and calculate
- **Weakness**: Hard to choose good set of categories to use; set of categories may lead some people to be overcounted or missed; accuracy of responses is unknown

Sources of problems related to hidden populations

In addition to the issues involved with estimating social network size, there are reasons to be concerned about the responses to these questions:

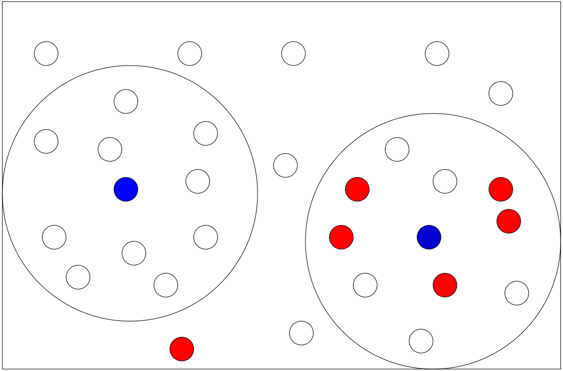
- How many people do you know who are drug injectors?
- How many people do you know who are sex workers?
- How many people do you know who are men who have sex with men?

Known problems: Barrier effects (i.e., non-random mixing)

Problem: hidden population is unevenly distributed in the population

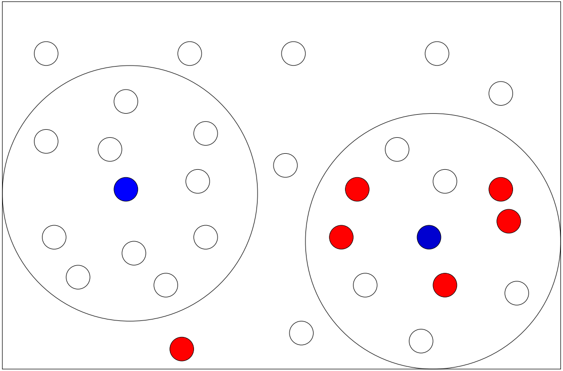
Known problems: Barrier effects (i.e., non-random mixing)

Problem: hidden population is unevenly distributed in the population



Known problems: Barrier effects (i.e., non-random mixing)

Problem: hidden population is unevenly distributed in the population



Consequence: Increase variance, but probably no effect on bias if sampling frame is complete

Known problems: Accuracy of responses

Problem: respondents might unable or unwilling to answer these questions accurately

- Transmission error (“masking”)
- Recall error
- Social desirability bias

Known problems: Accuracy of responses

Problem: respondents might unable or unwilling to answer these questions accurately

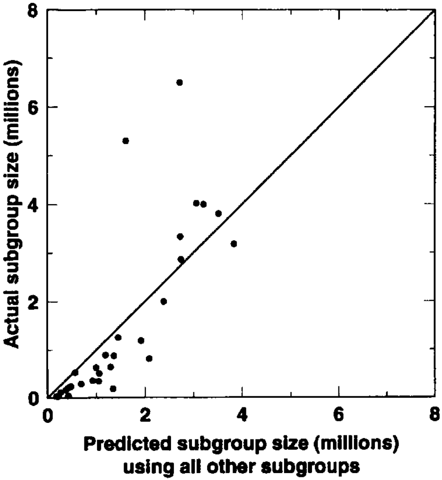
- Transmission error ("masking")
- Recall error
- Social desirability bias

Consequence: Could lead to under-estimate or over-estimate

Advantages of the network scale-up method

- Does not require contact with the hidden population
- Can be embedded into a nationally representative survey and takes about 5-10 minutes
- Possible to standardize across cities and countries
- Produces estimates for the sizes of many hidden populations at the same time
- Statistical methods are potentially improvable
- If back-estimation method is used for estimating social network size “reality checks” are possible

Reality checks



Source: Killworth et al, 1998

Limitations of the network scale-up method

- If people who know members of the in hidden populations are underrepresented in sample, estimates could be way too low
- Method has been mostly used in the United States
- We don't yet know exactly what affects the variance of the estimates so we don't know when the estimates will be precise enough to be useful
- We don't yet have a good procedure for putting confidence intervals around estimates
- If no correction is made for transmission error ("masking"), estimates could be way too low

Network scale-up study in Curitiba, Brazil

Two-part study to estimate the number of heavy drug users in Curitiba

- scale-up survey given to a random sample of the population ($n \approx 500$)
- transmission error survey given to a sample heavy drug users ($n \approx 300$)

Results will be combined to produce size estimates that will be compared to estimates using other methods.

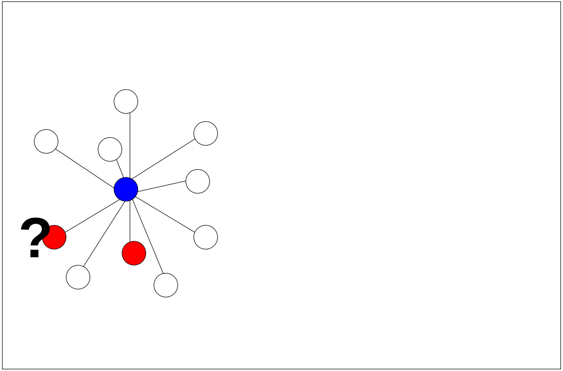
Funding: UNAIDS and Brazilian Ministry of Health

Know problems: Transmission error (i.e., masking)

Problem: Respondent may know someone who is an IDU, but not know they are an IDU

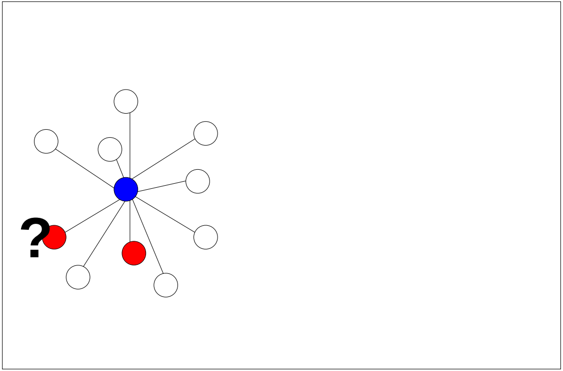
Know problems: Transmission error (i.e., masking)

Problem: Respondent may know someone who is an IDU, but not know they are an IDU



Know problems: Transmission error (i.e., masking)

Problem: Respondent may know someone who is an IDU, but not know they are an IDU



Consequence: Possibly a substantial underestimate of the size of the hidden population

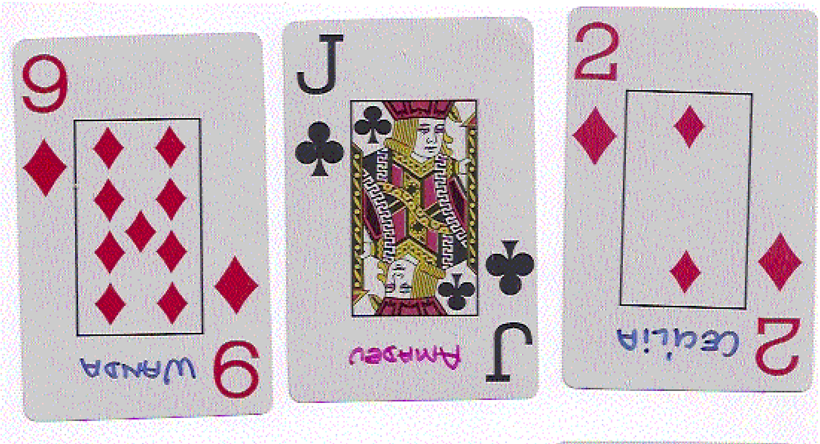
Transmission error study

We want to estimate the probability that a drug user's acquaintance knows that they are a drug user.

In other words, imagine my list of acquaintances:
1) Amanda
2) Adam B.
3) Adam S.
. . .
120) Zsuszi
What percentage of these people know that I am drug user?

Transmission error study

Respondent's don't have list of acquaintances that we can sample from so we will use a variation of the technique of McCarty et al. (1997). These interview begins when a deck of 24 playing cards is shuffled by the interviewer.



Transmission error study

Next a card is pulled from the deck and the respondent is asked

- How many people do you know named [SEBASTIAO]?

The respondent will pick up this many stones and place them on a board such as:

is a drug user knows I'm a drug user	is a drug user doesn't know I'm a drug user
is not a drug user knows I'm a drug user	is not a drug user doesn't know I'm a drug user

We will ask 12 male names and 12 female names.

Transmission error study

Given our (imperfect) information about the popularity of the names, we expect about 20 sampled alters per respondent.

These estimates will be noisy at the level of the individual, but for the adjustment we only need an average probability.

Results from transmission error study will be linked with demographics from the larger drug user study.

Thank you

Thank you

To learn more . . .

- Bernard, HR, et al (1989). “Estimating the Size of an Average Personal Network and of An Event Subpopualtion.” in *The Small World*, edited by M. Kochen. 159-175.
- Bernard, HR, et al (1991). “Estimating the Size of An Average Personal Network and of An Event Subpopulation: Some Empirical Results.” *Social Science Research*, 20:109-121.
- Killworth, PD, et al (1998). “Estimation of Seroprevlance, Rape, and Homelessness in the United States Using a Social Network Approach.” *Evaluation Review*, 22:289-308.
- McCarty, C, et al (2001). “Comparing Two Methods for Estimating Personal Network Size.” *Human Organization*, 60:28-39.
- Killworth PD, et al (2003). “Two Interpretations of Reports of Knowledge of Subpopulation Sizes.” *Social Networks*, 25:141-160.
- Killworth, PD, et al (2006). “Investigating the Variation of Personal Network Size Under Unknown Error Conditions.” *Sociological Methods & Research*, 35:84-112.
- Shelley, GA, et al (2006). “Who Knows Your HIV Status II? Information Propagation Within Social Networks of Seropositive People.” *Human Organization*, 65:430-444.
- Kadushin, C, et al (2006). “Scale-Up Methods As Applied to Estimates of Heroin Use.” *Journal of Drug Issues*, 36:417-440.
- Zheng, T, et al (2006). “How Many People do you Know in Prison?: Estimating Overdispersion in Count Data to Estimate Social Structure in Networks.” *Journal of the American Statistical Association*, 101:409-423.
- Snidero, S, et al (2008). “Question order and interviewer effects in CATI scale-up surveys.” *Under review*.
- Snidero, S, et al (2008). “Scale-up estimators in CATI surveys for estimating the number of foreign body injuries in the aero-digestive tract in children.” *Under review*.
- McCormick, T et al (2008). “How Many People do you Know?: Estimating Personal Network Size.” *Journal of the American Statistical Association*, in press.
- UNAIDS (2003). Estimating The Size of Population at Risk for HIV. UNAIDS.