How to conduct a network scale-up survey

Christopher McCarty and H. Russell Bernard
University of Florida

February, 2009

© 2009 Christopher McCarty and H. Russell Bernard

Network scale-up begins like most surveys

- Define respondent population
- Choose sample frame
- Choose survey mode
- Choose sample size
- Design questionnaire (This is the part that's different)

Selecting respondent population

- Respondent population is not the same as the population to be estimated (target population)
  - U.S. respondents to estimate homeless population
  - Urban population to estimate heroin users
- You must know the size of the respondent population
- Do transmission and barrier errors suggest using a respondent population with more ties to target population?
- This opportunity to do this research in multiple countries could help solve this problem

Choose sample frame

- The sample frame represents the respondent population
- For our work we used random digit dial telephone numbers
- For face-to-face a general population survey may rely on census or voter registration data

Choosing mode

- There are five survey modes
  - Face-to-face
  - Telephone (this is what we used)
  - Mail
  - Drop and collect
  - Web
- There is a large literature on mode effects in surveys
- For the populations of interest to UNAIDS a face-to-face or mixed mode makes sense

Choose sample size

- As with any survey, the sample size should be based on expected margins of error
- For this survey we have margins of error associated with network size
- Although estimates of network size are remarkably reliable, they have large standard deviations
- Our data suggest that a survey of 400 respondents would generate a margin of error of ±26 alters
- A survey of 1,000 would generate a margin of error of ±16 alters
- Keep in mind these are based on variance for U.S. respondents

Design questionnaire

- Network scale-up questionnaire has three parts

  1. Demographics used to estimate bias
  2. Question to estimate the number of alters respondents knows in the target population
  3. Questions to estimate network size (c)

- Steps 2 and 3 require a boundary definition of who is counted as a network alter

Alter boundary

- Definition of who is an alter can have enormous effects on the estimate
- Defining the alter boundary as 12 months will generate different network sizes than a boundary of two years
- Our definition:
  - You know them and they know you by sight or by name. You have had some form of contact with them in the past two years and you could contact them if you had to
  - Question: Should respondents be instructed to exclude those met on networking sites such as Facebook?

There are two ways to estimate c

- Scaling from known populations
- The summation method

Using known populations

- Select a set of known populations, the more the better
- Populations should vary in size and type
  - Limiting the study to populations related to health conditions, although plentiful, may introduce barrier error
  - Using only large populations (such as men or people over age 65) introduces a lot of estimation error
  - Using only small populations introduces error from very few hits
  - Known populations should be within .1% to 4% of population (this may change as we learn more)
- The demographic characteristics of the known populations should match as closely as possible the demographic characteristics of the population upon which the known estimates are based
- Populations are often related to transmission and barrier effects
- In the past we assumed that by using populations of multiple size and type these effects are cancelled out

Examples of populations we used

- In the U.S. there are a variety of sources for known populations:
  - The U.S. Statistical Abstract
  - The U.S. Census
  - The FBI Crime Statistics
- Ideally collection of sub-population data will be recurring so that they can be used in subsequent years
- It is important that the data all reflect the same year (be aware that some population data lags)
- Known populations are very susceptible to transmission and barrier error

Relationship between number known and demographic characteristics

Population	State	Sex	Race	Age	Education	Marital status	Work status	Religion	Political Party
Native Americans	•		•			•			
Gave birth in past 12 months			•	•	•	•	•	•	
Adopted a child in past year					•			•	
Widow(er) under 65 years				•		•	•	•	•
On kidney dialysis								•	
Postal worker		•	•					•	
Commercial pilot		•		•	•				•
Member of Jaycees	•	•			•	•	•	•	•
Diabetic					•	•	•	•	
Opened a business in year				•	•	•	•		
Have a twin brother or sister	•	•	•	•	•		•	•	•
Licensed gun dealer		•							
Came down with AIDS				•	•	•	•		
Males in prison		•	•	•			•		
Homicide victim in past year	•		•		•	•			
Suicide in past year			•		•				
Died in wreck in past year	•			•			•	•	
Women raped in past year				•	•				
Homeless			•			•			
HIV positive				•	•	•	•		

We experimented with names

- Census provides estimates of both first names and last names
- We experimented with both types and found problems with each
- The advantage of names is that they vary in size and are typically ascribed
- Countries and cultures vary in the way they use names
- They are prone to barrier error

Relationship between number known and demographic characteristics

Population	State	Sex	Race	Age	Education	Marital status	Work status	Religion	Political Party
Michael		•		•	•	•	•	•	•
Christina		•		•	•	•	•	•	•
Christopher		•		•	•	•	•		•
Jacqueline			•	•	•		•	•	
James		•	•		•	•	•	•	•
Jennifer			•	•	•	•	•	•	•
Anthony	•	•	•		•		•	•	•
Kimberly	•		•	•	•	•	•	•	•
Robert		•		•	•	•	•	•	•
Stephanie				•	•	•	•	•	•
David		•		•	•	•	•	•	•
Nicole				•	•	•	•		

Summation method

- We can estimate network size (c) directly by asking respondents to tell us how many people they know
- This is an unreasonable task unless it is broken into reasonable subtasks
- We use culturally relevant categories of relation types that are mutually exclusive and exhaustive
- These are small enough that respondents can estimate them reliably

Relation categories we used

- **Immediate family**
- **Other birth family**
- **Family of spouse or significant other**
- **Co-workers**
- **People at work but don't work with directly**
- **Best friends/confidantes**
- **People know through hobbies/recreation**
- **People from religious organization**
- **People from other organization**
- **School relations**
- **Neighbors**
- **Just friends**
- **People known through others**
- **Childhood relations**
- **People who provide a service**
- **Other**

Developing a protocol for discovering summation categories

- We assume that relation categories used to elicit estimates will be culturally relative
  - Different languages will require their own category names
  - The way people maintain people in their mind will almost certainly vary by culture
- Further research is needed to determine the best protocol for discovering these categories
- Summation categories must be mutually exclusive, exhaustive and small enough that respondents count rather than estimate

Approaches we are studying

- Our current categories emerged from a previous study about the ways people know each other
- This is not ideally suited to this study
- We are exploring using cultural consensus analysis or personal network structure to quickly develop these categories
- An empirical approach is to start with very large culturally relevant categories and use alter characteristics to split them when they are too large

Estimates of network size from two methods (scaling from known and summation) are very close

- Scaling from known populations
  - 290.8 (SD 264.4)
- Summation method
  - 290.7 (SD 258.8)
- We checked in multiple ways to see whether this was an artifact of the method
- It wasn't

Advantages of the summation method

- It is quicker, taking about half the time or less than estimating from known sub-populations
- It should not be subject to transmission or barrier error for estimates of network size
- It does not require finding known populations, which could be a problem in some countries

Disadvantages of summation method

- It cannot be verified statistically
- It may be easy for respondents to double count network alters as they are multiplex relations (such as co-worker and social contact)
- Network size calculated from scaling known populations can be checked by back-estimating each known with the other knowns

Modeling issues

- At this point in our work we are convinced that our estimates of network size are *relatively* reliable, but not *absolutely* reliable
- If my network is 300 then I am confident it is half as large as that of someone with a network of size 600
- I am not confident that the network size is actually 300
- This compromises our ability to estimate the absolute size of a population
- Again, the opportunity to replicate this method may yield solutions

How to generate scale-up estimates

- There are two steps
  - Estimate network size c
  - Use c with respondents' estimates of unknown populations to scale-up to the size of the unknown in the population
- We will look at these steps separately

Step 1: Estimating c using summation method

- With the summation method you add up the estimates form each relation category to get a c value for each respondent
- The c used in the formula will be the average of all those c values from each respondent

Step 1: Estimating c using known populations

- This procedure requires three parameters
- t=the size of the population to which you are scaling up (this is the same for each respondent)
- e=the sum of all the known populations you are using in the survey (this is the same for each respondent)
- m=the sum of all the reported known subpopulation sizes for each respondent
- c for each respondent is (m*t)/e
- The c used in the formula will be the average of all those c values from each respondent

Step 2: Applying c

- This step also requires three parameters
- t=the size of the population to which you are scaling up (this is the same for each respondent)
- c=the average c value, either from the scale-up or the summation method
- m=the average of all respondents' estimates of the number of people they know in the unknown subpopulation
- The formula to estimate the size of the unknown subpopulation e=(m/c)*t