Methods for Estimating Hard to Count Populations

Adrian E. Raftery
University of Washington
http://www.stat.washington.edu/raftery

NAS Committee on Best Practices in Assessing Mortality and Significant Morbidity Following Large-Scale Disasters Webinar

February 11, 2020

Outline

- Capture-recapture methods for combining multiple fragmentary data sources[1]
- Respondent-driven sampling[2] [3]
- Network scale-up method[4] [to be presented by Tyler McCormick]

[1] Bao, Raftery, Reddy (2015, *Statistics and its Interface*)
[2] Baraff, McCormick, Raftery (2016, PNAS)
[3] Green, Raftery, McCormick (2020, *Biometrika*)
[4] Maltiel, Raftery, McCormick, Baraff (2015, *Annals of Applied Statistics*)

Uncertainty About Hidden Population Size

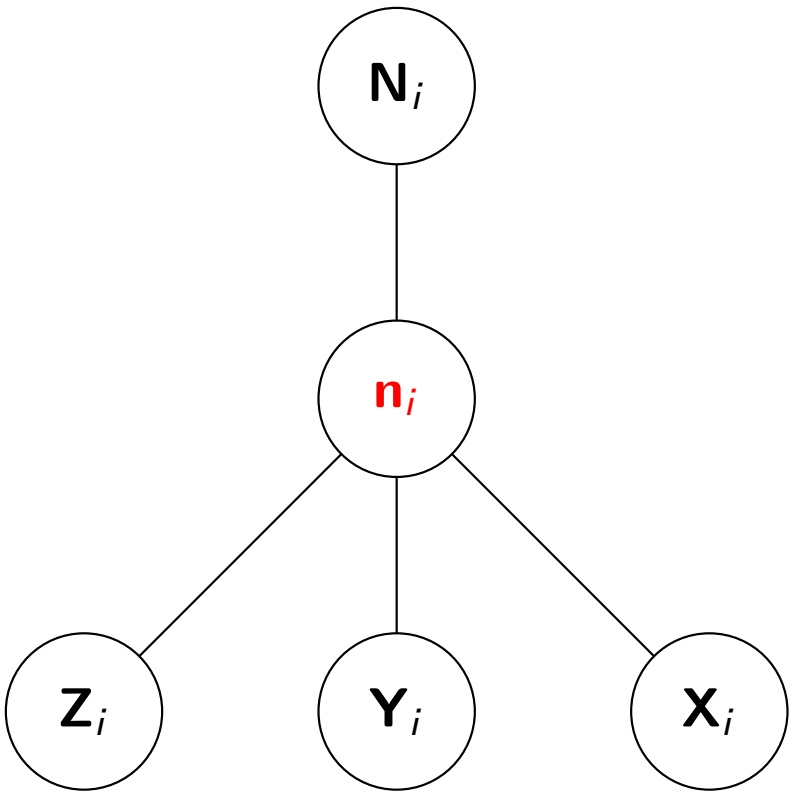
- Intrinsically hard to assess because of confounding between population size and probability of being counted:
  - It's hard to distinguish between low population size (close to the number seen) with high probability of being counted, and
  - high population size with low probability of being counted
  - $\Longrightarrow$ uncertainty about N can be large
  - Especially for estimation from one data source
  - Prior or external information can be useful, especially:
    - information about detection probabilities
    - a (possibly soft) upper bound on N when detection probabilities are small
- Bayesian approach:
  - can combine information of different types
  - can deal with missing data
  - allows one to use data from other periods, regions, countries to inform the prior distribution

Capture-Recapture Estimation

Data Sources for Size Estimation of Intravenous Drug Users in Bangladesh, 2004

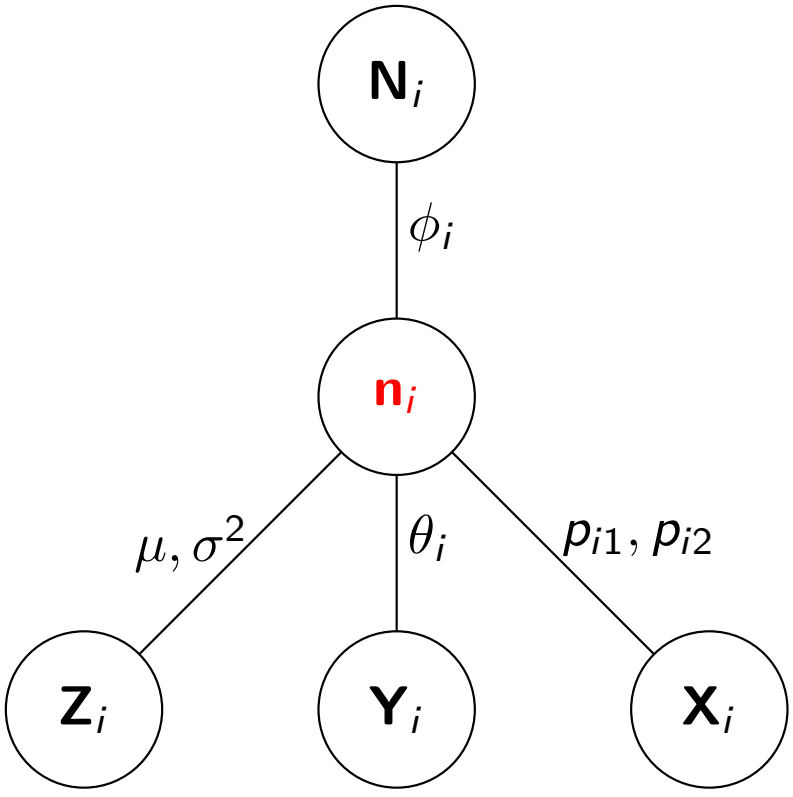
Notation	Data Source	Locations	Year
N_i	adult male population size	64 districts	2003
Y_i	NASROB mapping data	24 districts	2001
$X_{i10} + X_{i11}$	NEP intervention	3 districts	2001
$X_{i01} + X_{i11}$	BSS survey	4 districts	2002
X_{i11}	reported NEP enrollment in BSS	2 districts	2002
Z_i	CARE RSA estimate	14 districts	2003

Data Structure



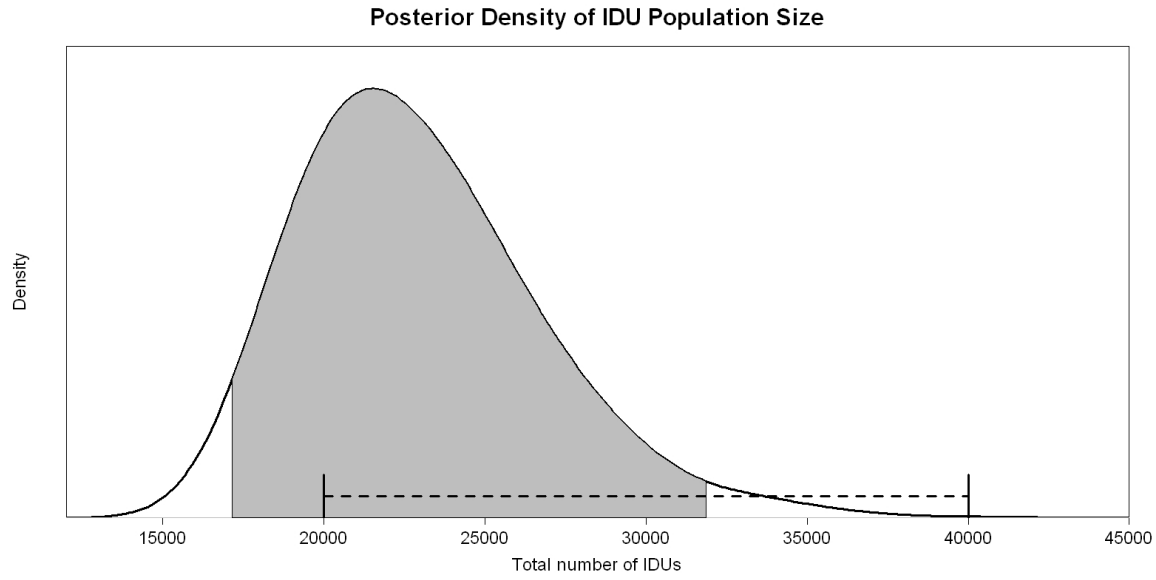
- For the ith district:
  - N_i = total population
  - n_i = number of IDUs
  - Z_i = estimate of n_i from other sources
  - Y_i = the number of IDUs in survey or intervention
  - $X_i = (X_{i10}, X_{i01}, X_{i11})$ = data from two lists from multiplier method or capture-recapture

Bayesian Hierarchical Model



- For the ith district:
  - $n_i | N_i, \phi_i \sim \text{Binomial}(N_i, \phi_i)$
  - $\log(\frac{z_i}{n_i}) \sim \text{N}(\mu, \sigma^2)$
  - $y_i | n_i, \theta_i \sim \text{Binomial}(n_i, \theta_i)$
  - $x_{i11} + x_{i10} | n_i, p_{i1} \sim \text{Binomial}(n_i, p_{i1})$
  - $x_{i11} + x_{i01} | n_i, p_{i2} \sim \text{Binomial}(n_i, p_{i2})$
  - $x_{i00} | x_{i11}, x_{i10}, x_{i01} \sim \text{NegBinomial}(x_{i11} + x_{i10} + x_{i01}, 1 - (1 - p_{i1}) \times (1 - p_{i2}))$

Population Size Estimate



- Posterior median 23,500
- Posterior interval 18,000 $\sim$ 33,700
- Bangladesh technical group estimate: 20,000 $\sim$ 40,000

Respondent-Driven Sampling

Introduction

Respondent-Driven Sampling (RDS):

- Problem - no sampling frame for hard-to-reach or hidden populations (IDU, FSW, MSM, etc.)
- Developed by Heckathorn (1997, 2002) to add statistical rigor to snowball (chain-referral) sampling
  - known biases due to homophily and centrality
- Start with small number of initial respondents (seeds)
  - usually convenience sample
- Seeds given coupons to distribute to peers
- When coupons redeemed, respondents are given new coupons creating chain of recruitment
- Additional information gathered
  - coupon number - who recruited whom
  - degree of each respondent - how many could have recruited

Estimation

Volz-Heckathorn Estimator (2008):

- To estimate the mean μ of a trait

$$\hat{\mu}_{VH} = \frac{\sum_{i=1}^{n} \frac{x_i}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}},$$

  where x_i = value of the trait in individual i and
  d_i = degree of individual i
- Hansen-Hurwitz (H-H) type estimator with selection probabilities proportional to degree
- Unbiased if Markov chain model holds

Problem: How do we estimate $\text{Var}(\hat{\mu}_{VH})$?

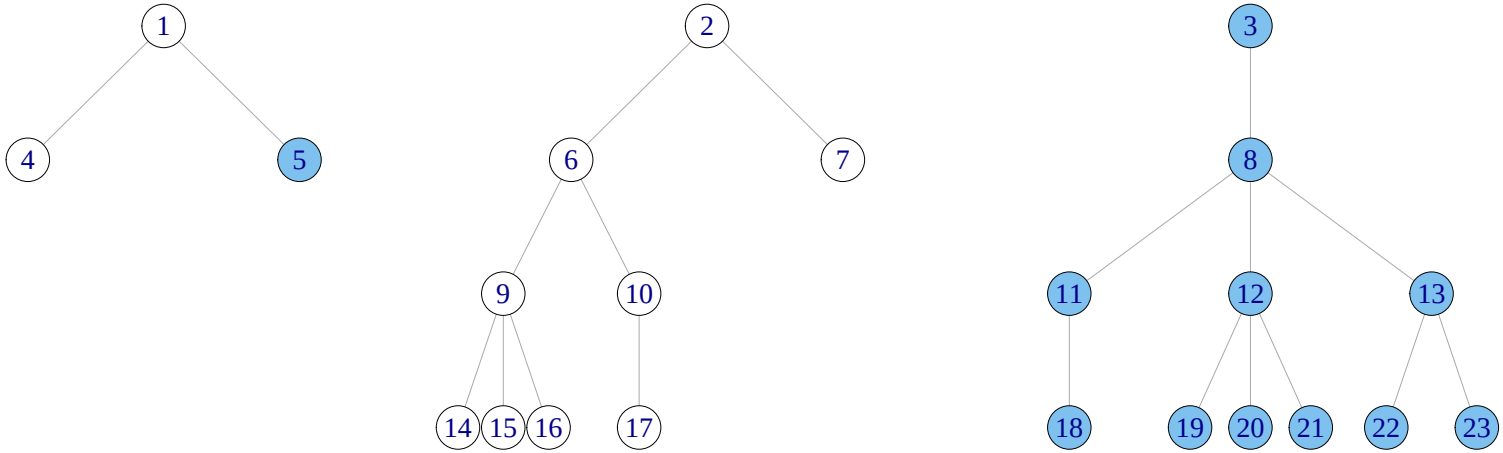
Tree Bootstrap

My idea:

- Ignore traits and focus on the tree structure of the RDS
- Multilevel bootstrap
  1. Resample seeds with replacement
  2. From each seed resample recruits with replacement
  3. From each recruit resample further recruits with replacement
  4. Repeat 3 until no recruits are available
- Assumes the parts of the network unseen "look" like the parts of the network seen
- Benefits
  - Correlation structure of entire tree preserved, not just first-order transitions
  - One bootstrap sample for all traits

Tree Bootstrap

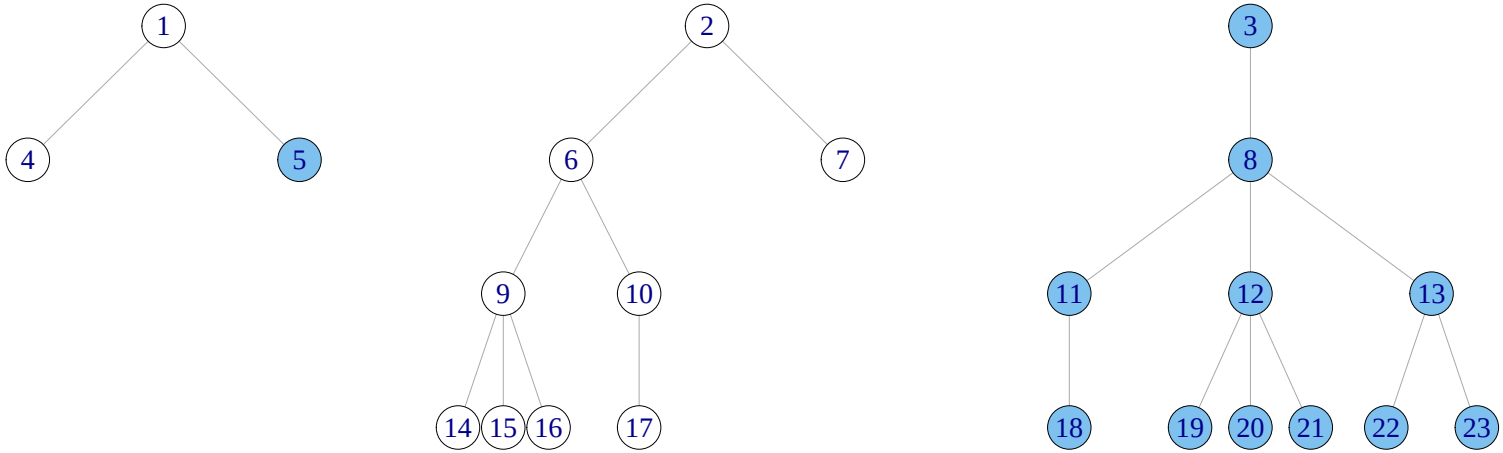
Sample:



Bootstrap:

Tree Bootstrap

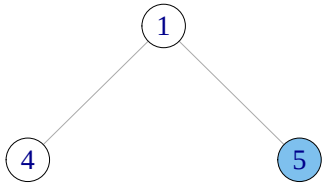
Sample:

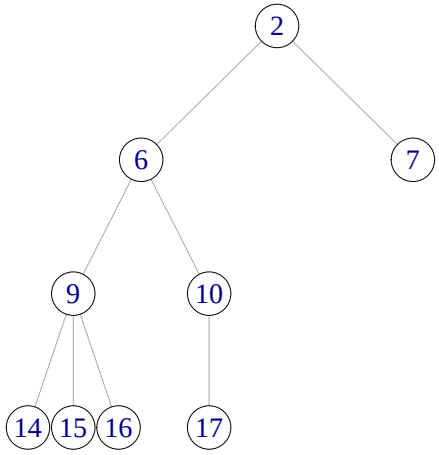


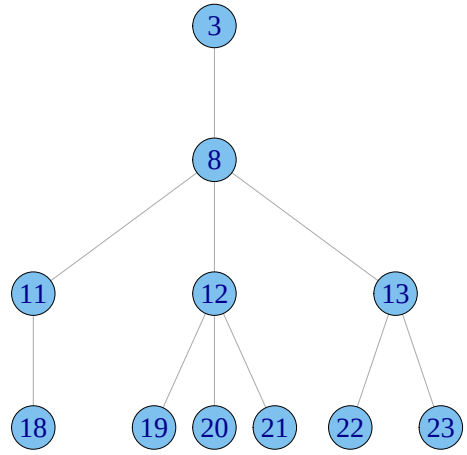
Bootstrap:

Tree Bootstrap

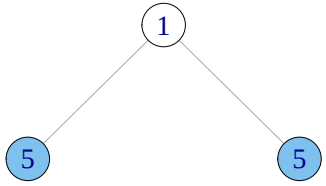
Sample:

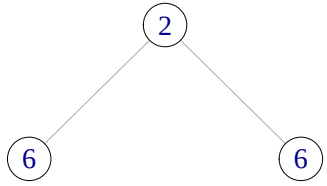


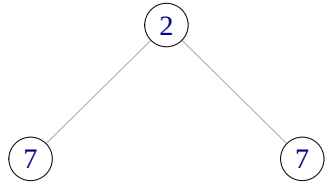




Bootstrap:

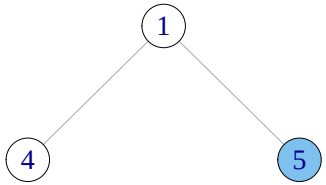


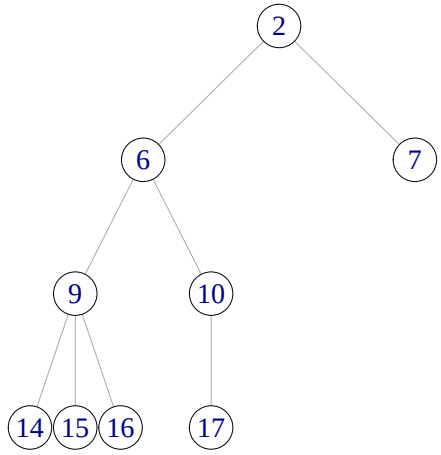


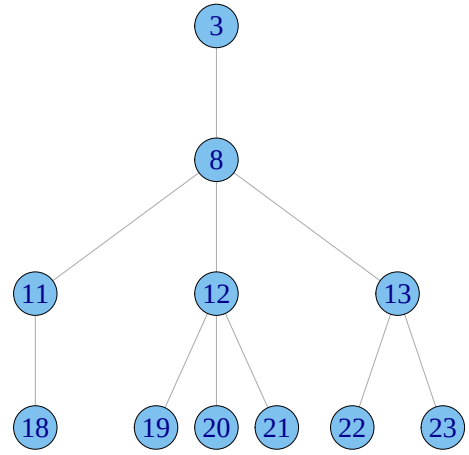


Tree Bootstrap

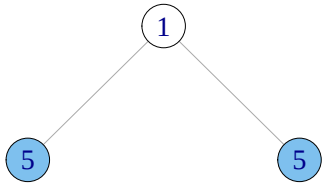
Sample:

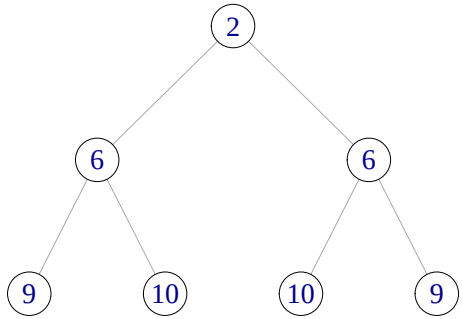


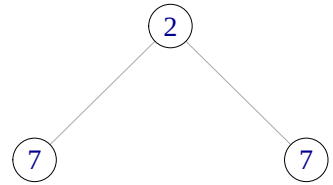




Bootstrap:

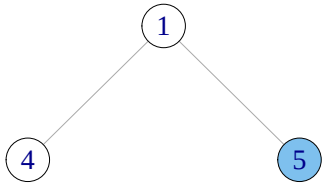


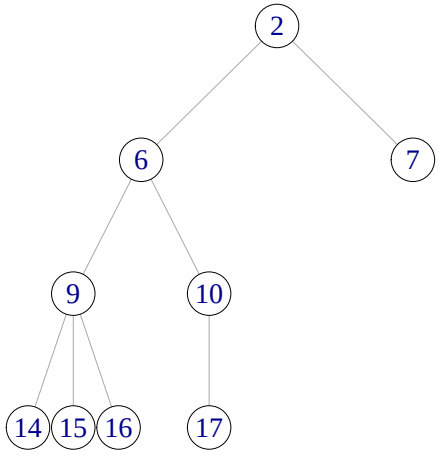


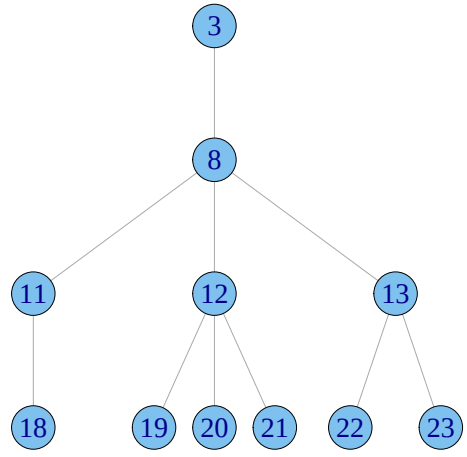


Tree Bootstrap

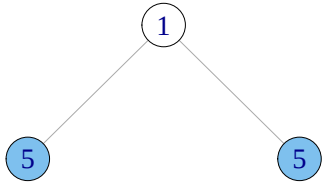
Sample:

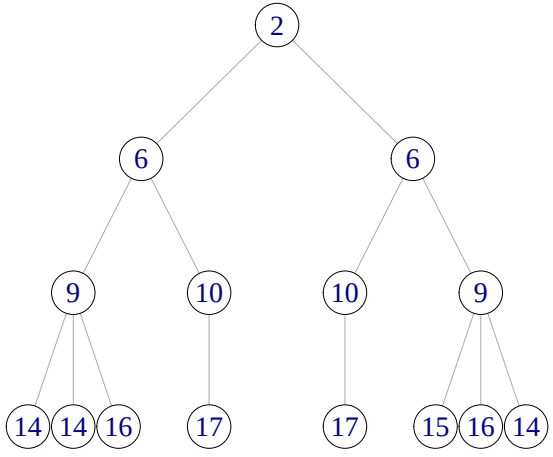


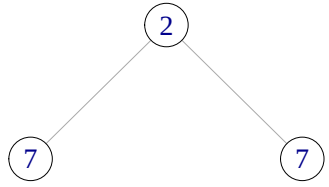




Bootstrap:







- Tree bootstrap performed better than competing methods in a simulation study[5]
- Tree bootstrap estimate of variance is asymptotically correct[6]

[5] Baraff, McCormick, Raftery (2016, *PNAS*)
[6] Green, McCormick, Raftery (2020, *Biometrika*)