Estimating the Size of Hidden Populations using Respondent-Driven Sampling Data

Mark S. Handcock  
Department of Statistics  
University of California - Los Angeles

Krista J. Gile  
Department of Mathematics  
University of Massachusetts - Amherst

Corinne M. Mar  
Center for Studies in Demography and Ecology  
University of Washington

Supported by NIH grants 1R21HD063000 and 5R21HD075714-02, NSF awards MMS-0851555, SES-1230081 and MMS-1357619 and the DoD ONR MURI award N00014-08-1-1015.

Working Papers available at  
http://www.stat.ucla.edu/~handcock  
http://arXiv.org

UNAIDS Reference Group Consultation, UMass, June 9-10, 2014
Hard-to-Reach Population Methods Research Group

- Isabelle Beaudry, UMass
- Ian E. Fellows, Fellows Statistics
- Krista J. Gile, UMass
- Mark S. Handcock, UCLA
- Lisa G. Johnston, Tulane University, UCSF
- Corinne M. Mar, University of Washington

http://hpmrg.org
Inferential approaches

The key is the modeling of the sampling process

- Salganik and Heckathorn (2004): Markov chain model over classes
- Volz and Heckathorn (2008): Markov chain model over people
Successive Sampling (SS)

Consider the following *successive sampling* (SS) or *probability proportional to size without replacement* (PPSWOR) sampling procedure:

- Begin with a population of \( N \) units, denoted by indices 1 \ldots N with varying sizes represented by \( d_1, d_2, \ldots d_N \).
- Let \( G_1, \ldots, G_N \) be the indices of the successively sampled people.
- Sample the first unit from the full population \( \{1 \ldots N\} \) with probability proportional to size \( d_i \). Assign the index of this unit to the random variable \( G_i \).
- Select each subsequent unit with probability proportional to size *from among the remaining units*, such that

\[
P(G_i = k | G_1 \ldots G_{i-1}) = \begin{cases} \frac{d_k}{\sum_{j \notin \{G_1 \ldots G_{i-1}\}} d_j} & k \notin \{G_1 \ldots G_{i-1}\} \\ 0 & \text{else} \end{cases}
\]
Gile (2011) argues that RDS can be approximated by Successive Sampling under a *configuration model* for the network:

- Node $i$ has given degree, $d_i$, consider $d_i$ edge-ends.
- Pairs of edge-ends matched up at random
- This is a *configuration model*

- Suppose $G_1, G_2, \ldots G_k$ by Successive Sampling according to $d$.
- Then if the network is unknown, but known to be a configuration model, tracing a link from $G_k$ will select nodes according to Successive Sampling.
Is there information in RDS data about population size?

Idea:

- Under Successive Sampling, “larger” units typically sampled earlier
- Early sample: lots of “big” units, few “small”
- Later sample: fewer “big”, more “small”
- *No change implies the population is not much depleted*
- *Big change implies population very depleted*

This can be quantified to estimate \( N \)!

*Note: the information about \( N \) is in the ordered sample pattern!*
Modeling the sampling process for non-ignorable sampling

- RDS is not ignorable: \( P(G|D_{obs}, D_{unobs}) \neq P(G|D_{obs}) \)
- Information about \( N \) is in the sequence of observations.
- Make inference from joint model for sample sequence and unit sizes.
Inferential Approach

Observed data:

\[ D_{obs} = \text{Unit sizes (degrees) of observed units in order of observation} \]

Goal:

\[ P(N|D_{obs}) \]

(posterior distribution of \( N \) given the data)

Parameters:

\( N = \text{Population Size} \)

\( \eta = \text{Parameter of distribution of unit sizes} \).
Inferential Approach

\[ P(N|D_{obs}) \propto \int P(D_{obs}|N, \eta)P(\eta, N)d\eta \]

(independent priors)

\[ = \int P(D_{obs}|N, \eta)P(\eta)P(N)d\eta \]

\[ = \int \text{(likelihood)} \ (\text{prior for } \eta) \ (\text{prior for } N)d\eta \]

\[ = \int P(D_{obs}|G, U, D, \eta)P(G|U, \eta)P(U|N, \eta)P(\eta)P(N)d\eta \]

\[ = \int \text{P(samp given degrees)} \ P(\text{degrees}) \ (\text{prior } \eta) \ (\text{prior } N)d\eta \]
Parametric model for the degrees:

\[ d_i \sim f(\cdot | \eta) \]

with support \( d = 0, 1, \ldots \), and parameter \( \eta \).
Models for Degrees

Extensive papers by Handcock and Jones. To specify $f(\cdot | \eta)$. We can consider:

1. Poisson
2. Negative binomial. This allows Gamma over-dispersion over Poisson.
3. Yule, Waring. This allows power-law over-dispersion over Poisson.
4. Poisson-log-normal. This allows log-normal over-dispersion over Poisson. It is more than the Negative Binomial but less than the power-law models.
5. Conway-Maxwell-Poisson distribution. This allows both under-dispersion and over dispersion with a single additional parameter over a Poisson.
6. Non-parametric lower tails: To allow for poor fit in the lower degrees.

These are all coded up in the CRAN degreenet package and/or the size package.
Prior for the degree distribution model

Each degree distribution model parametrized with mean and standard deviation.

\[ \eta = g(\mu, \sigma) \]

\[ \mu | \sigma \sim N(\mu_0, \sigma_0/df_{\text{mean}}) \quad \sigma \sim \text{Inv}\chi(\sigma_0; df_{\text{sigma}}) \]

Use diffuse default prior on degree model parameters (equivalent sample size \( df_\mu = 1 \) and \( df_\sigma = 5 \)).
Many possibilities

- The data truncates the prior below the sample size.
- Uniform prior is improper
- Natural parametric models (e.g., Negative Binomial, Poisson-log-normal, Conway-Maxwell-Poisson).
- Natural parametric models too thin in the tails
- Instead: specify prior knowledge about the sample proportion (i.e. $n/N$).
Prior for Population Size $N$

- Simple prior: uniform on $n/N$.
- Gives closed form prior on $N$ with infinite mean (median = $2n$)
- Generalize to $n/N \sim \text{Beta}(\alpha, \beta)$

The density on $N$ (considered continuous) is:

$$\pi(N) = \frac{\beta n(N - n)^{\beta-1}}{N^{\alpha+\beta}} \quad \text{for} \quad N > n.$$ 

- The distribution has tail behavior $O(1/N^{\alpha+1})$.
- Elicit median or mode from field researchers and translate to $\beta$ and/or $\alpha$. 

Figure: Three example prior distributions for the population size ($N$). They correspond to $\alpha = 1$ and $\beta = 1.55, 1.16$ and $3$. 
Likelihood: Notation

Let:
\[ D_{obs} = (D_1, \ldots, D_n) \] be the random ordered observed degrees (ordered for notation)
\[ D_{unobs} = (D_{n+1}, \ldots, D_N) \] be the unordered random unobserved degrees
Let \( d_{obs} = (d_1, \ldots, d_n) \) and \( d_{unobs} = (d_{n+1}, \ldots, d_N) \) be their realized values.
Let \( G = (G_1, \ldots, G_n) \) be the random indices of the ordered sample and \( g_{obs} \) be the observed sequence.
Likelihood

\[ P(D_{\text{obs}}|N, \eta) \]
\[
= \sum_{d} \sum_{g} p(D_{\text{obs}} = d_{\text{obs}}|G = g, D = d, \eta)p(G = g|D = d, \eta)p(D = d|\eta)
\]
\[
= \frac{N!}{(N - n)!} \sum_{d \in \mathcal{D}U(d_{\text{obs}})} p(G = (1, \ldots, n)|D = d) \prod_{j=1}^{N} f(d_j|\eta)
\]

where \( \mathcal{D}U(d_{\text{obs}}) \) is the set of possible \( d_{\text{unobs}} \) given \( d_{\text{obs}} \).

\[ P(G|D = d, N, \eta) = \prod_{k=1}^{n} \frac{d_k}{\lambda_k} \]

where

\[ \lambda_k = \sum_{i=k}^{N} d_i = \sum_{i=k}^{n} d_i + \sum_{i=n+1}^{N} d_i \quad k = 1, \ldots, n \]

depends on both \( d_{\text{obs}} \) and \( d_{\text{unobs}} \).

\[ L[N, \eta|D_{\text{obs}}, G] \propto \frac{N!}{(N - n)!} \sum_{d_{\text{unobs}} \in \mathcal{D}U(d_{\text{obs}})} \prod_{k=1}^{n} \frac{d_k}{\lambda_j} \cdot \prod_{j=1}^{N} f(d_j|\eta) \]
Inference

- Likelihood can be maximized
- Can combine with priors to compute posterior
- Note computational complexities based on sum over $N - n$ embedded sums over infinite spaces.
Example: $N=1000$, homophily=2, diff. activity=3
Application: Estimating the numbers of those most at risk for HIV in Cities in El Salvador

- Surveillance surveys in El Salvador
- Focus on high-risk groups: female sex workers (FSW)
- RDS study of size \( n = 184 \) in 2010.
Figure: Graphical representation of the recruitment tree for the sampling of female sex workers in Sonsonate, El Salvador in 2010. The nodes are the respondents and the wave number increases as you go down the page. The node gray scale is proportional to the network size reported by the worker, with white being degree one and black the maximum degree.
Figure: Posterior distribution for the number of female sex workers in Sonsonate based on three prior distributions for the population size: flat, matching the midpoint UNAIDS estimate, and interval-matching the UNAIDS estimate. The prior is dashed. The red mark is at the posterior median. The green mark is at the posterior mean. The blue lines are at the lower and upper bounds of the 95% highest-probability-density-interval. The purple lines demark the lower and upper UNAIDS guidelines.
Simulation Study

Simulate Population

- 1000, 835, 715, 625, 555, or 525 nodes
- 20% “Infected”

Simulate Social Network (from ERGM, using statnet)

- Mean degree 7
- Homophily on Infection: \( \alpha = \frac{\mathbb{E}(\text{# infected to infected tie})}{\mathbb{E}_{R=0}(\text{# infected to infected tie})} = 5 \) (or other)
- Differential Activity: \( \omega = \frac{\text{mean degree infected}}{\text{mean degree uninfected}} = 1 \) (or other)

Simulate Respondent-Driven Sample

- 500 total samples
- 10 seeds, chosen proportional to degree
- 2 coupons each
- Coupons at random to relations
- Sample without replacement

Blue parameters varied in study.
Evaluating Performance: Frequentist properties of Bayesian method

- Point estimates: are they about right on average?
- Using the Bayesian framework, use probability intervals for the population size (Highest Posterior Density Credible Intervals - CI’s)
- Compare Frequentist properties: CI width and coverage rates
**Figure:** Spread of central 95% of simulated population size estimates (posterior means) for 5 population sizes for low, accurate, and high priors. Dots represent means. Estimates are represented as multiples of the true population size (red line at 1 indicates true population size). Numbers below the bars are coverage rates of 95% HPD intervals.
Miss-specification of the Network Structure

Figure: Spread of central 95% of simulated population size estimates (posterior means) for population size 1000 for low, accurate, and high priors, with varying levels of homophily ($\alpha$) and differential activity ($\omega$).
Discussion

- It is important to estimate the networked population size
- There is information on the population size implicit in the decreasing degrees of the sample nodes over time
- Using successive sampling model we can model the decrease
- We can incorporate prior information about the population size using the Bayesian framework
- We can incorporate other features of the population
- We can estimate population means (e.g., prevalence and counts)
- In the Bayesian framework, we can estimate uncertainty of the estimates in a natural way
Cautions

- The difference between the model with disease and without highlights the importance of the specification of the model for the degree distribution.
- The estimates depend on the prior distribution for population size.
- The estimates are biased because the successive sampling model is not perfect - and will be be increasingly misspecified as we get further from the configuration network.
- There is another important (and influential) tuning parameter: K - the truncation value for degrees.
- This approach is promising. It is designed to be combined with data from other methods (e.g., scale up) to provide the most accurate overall estimate.
- Fundamentally, RDS data typically does not contain much information about the population size. The Bayesian approach enables us to quantify this.
Prior for the population size

- translates to a closed form for the prior on $N$ which has infinite mean
- Generalize to $n/N \sim \text{Beta}(1, \beta)$

The density function on $N$ (considered as a continuous variable) is:

$$f(N|n) = \beta n(N - n)^{\beta - 1}/N^{\beta+1} \quad \text{for} \quad N > n$$

The distribution has tail behavior $\approx 1/N^2$. The mode of the prior is at $0.5n(\beta + 1)$ and the median is given by $n/(1 - (1/2)^{1/\beta})$. The median or mode can be elicited from field researchers and translated to $\beta$. A uniform distribution on the sample proportion corresponds to a median of twice the sample size.
Prior for population size

Prior mode = 1000

Prior for population size

Prior mode = 10000

truth=1000

population size

prior density
Example: $N=1000$, homophily=2, diff. activity=3

posterior for population size

Density

population size
200 RDS samples, prior.size.mode = truth
circle is mode, triangle is mean

Differential Activity Level
Homophily Ratio 1
Homophily Ratio 5
Because of differential activity the Conway-Maxwell-Poisson model class does not cover the bimodality of the degree distribution
Posterior for Population Size

truth = 1000  
mode = 763  
median = 801  
mean = 835
Miss-specification of the degree distribution shape

- Because of differential activity the Conway-Maxwell-Poisson model class does not cover the bimodality of the degree distribution.
- Solution: Model the degree distributions of the diseased from the non-diseased with separate Conway-Maxwell-Poisson models.
Posterior for Population Size

truth = 1000 mode = 988 median = 993 mean = 1049

Posterior for Mean Degree: true overall mean degree is 7

Posterior for s.d. degree
Modeling of other characteristics of the population

- As we model many population characteristics, including the disease status, we can compute estimates of them directly.
- Example: disease prevalence
Disease Prevalence REVISION 170
200 RDS samples, prior.size.mode = truth
Population size: red is 525, green is 715, and blue is 1000
Confidence intervals, design effects and standard errors

- Using the Bayesian framework, we can naturally compute the probability intervals for the population size and other characteristics
- Examples: CI coverage for the population size and prevalence
REVISION 170
200 RDS samples, prior.size.mode = truth
Coverage: proportion of samples whose 95% CI covered the true population size
True population size: red is 525, green is 715, blue is 1000

Differential Activity Level
Homophily Ratio 1 Homophily Ratio 5

REVISION 170
200 RDS samples, prior sampling fraction = truth
Coverage: proportion of samples whose 95% CI covered the true prevalence
True prevalence = 0.2
Application: The study of HIV/AIDS in San Francisco

- Surveillance surveys by San Francisco Department of Public Health
- Focus on African-American (AA) men-who-have-sex-with-men (MSM)
- RDS study of size $n = 256$ in 2009.
- Intensive study provides a population size estimate of 4439.
- Census data indicated 21518 AA men in San Francisco.
posterior for population size

population size

Density

0e+00

2e−05

4e−05

6e−05

8e−05

0 5000 10000 15000 20000

SF

median

0 5000 10000 15000 20000

posterior for population size

population size

Density

0e+00
posterior for the number of AA MSM with HIV
posterior for population size

Density

population size

SF mean
posterior for the number of AA MSM with HIV

HIV+ count
Density
SF mean
Discussion

▶ It is important to estimate the networked population size
▶ There is information on the population size implicit in the decreasing degrees of the sample nodes over time
▶ Using successive sampling model we can model the decrease
▶ We can incorporate prior information about the population size using the Bayesian framework
▶ In the Bayesian framework we can estimate uncertainty of the estimates in a natural way
Cautions

- The difference between the model with disease and without highlights the importance of the specification of the model for the degree distribution.
- The estimates depend on the prior distribution for population size.
- The estimates are biased because the successive sampling model is not perfect - and will be increasingly misspecified as we get further from the configuration network.
- There is another important (and influential) tuning parameter: K - the truncation value for degrees.
- This approach is promising. It is designed to be combined with data from other methods (e.g., scale up) to provide the most accurate overall estimate.
- Fundamentally, RDS data typically does not contain much information about the population size. The Bayesian approach enables us to quantify this.
Miss-specification of the degree distribution shape

- Because of differential activity the Conway-Maxwell-Poisson model class does not cover the bimodality of the degree distribution.
- Solution: Model the degree distributions of the diseased from the non-diseased with separate Conway-Maxwell-Poisson models.
Posterior for Mean Degree: true overall mean degree is 7

Posterior for s.d. degree

Population Size REVISION 170
200 RDS samples, prior.size.mode = truth
circle is mode, triangle is mean

Differential Activity Level
Homophily Ratio 1 Homophily Ratio 5
As we model many population characteristics, including the disease status, we can compute estimates of them directly.

Example: disease prevalence
Disease Prevalence

200 RDS samples, prior.size.mode = truth
Population size: red is 525, green is 715, and blue is 1000

Differential Activity Level

Homophily Ratio 1

Homophily Ratio 5
REVISION 170
200 RDS samples, prior sampling fraction = truth
Coverage: proportion of samples whose 95% CI covered the true prevalence
True prevalence = 0.2

Population Size

Homophily Ratio 1
Homophily Ratio 5
Comparison to the Gile SS estimator

- The SS estimator in Gile JASA (2011) requires $N$ known
- We use the posterior mode as a plug in estimate of $N$
Figure: Spread of central 95% of simulated prevalence estimates for population size 1000, with varying levels of homophily ($\alpha$) and differential activity ($\omega$). Solid lines represent prevalence estimates based on the posterior mean, dashed lines represent comparable estimates using the prior mean. Relative efficiency (MSE posterior/MSE prior) is given above each bar, and the coverage of nominal 95% confidence intervals is below each bar.