Respondent driven-sampling

Procedure to sample from hidden or hard-to-reach populations

Lucas Moschen

School of Applied Mathematics
Fundação Getulio Vargas

May 27, 2021

Table of Contents

Table of Contents

Hidden and hard-to-reach populations

- No sampling frame exists: size and boundaries of the population are unknown.
- Privacy concerns: stigmatized or illegal behavior.
- Fear of exposition or prosecution complicates the enumeration and learning about these populations.
- High logistic cost when the occurrence frequency is low.
- Examples: Heavy drug users, sex workers, homeless people, and men who have sex with men.

Existing sampling methods

- **Snowball** [Goodman, 1961]

From starting individuals, each subject provides a list of names of known individuals from the target population. The researcher invites this person to participate, who can agree or deny it.

- **Key informant** [Deaux and Callaghan, 1985]

Expert respondents are selected to answer about others' behavior. For instance, social workers, drug abuse counselors, official, etc.

- **Targeted** [Watters and Biernacki, 1989]

Field researchers build an ethnographic mapping of a target population, and recruit a number of individuals at sites identified by this map.

Problems with snowball sampling

- Inferences about the individuals depend on the initial sample.
- Bias towards individuals who are more cooperative and agree to participate.
- Bias because of masking, that is, protecting friends by not referring them.
- Individuals with more links may be oversampled.

Respondent-driven sampling

1. The researchers select a handful of individuals from a target population who serve as *seeds*.

2. Each participant receives a fixed number of *recruitment coupons* and invite members of their own social network to participate in exchange of a reward.

3. The sampling is without replacement.

4. If the individual accepts to participate, they answer a questionnaire and inform the network degree. One important point is that the recruiter doesn't say the name of the other members, reducing the mask effect.

System of incentives

Two different sources of theoretical incentive (dual incentive system):

- **Individual-sanction based control:** reward for participating in the research.
- **Group-mediated social control:** reward for recruiting peers. When social approval is important, it's more efficient and cheaper. Symbolic incentive is also important.

Table of Contents

Formal model

The RDS can be seen mathematically in two different approaches.

- **Stochastic process** [Heckathorn, 1997]

Each recruiter's social characteristics affect the characteristics of the recruits. There are a limited number of states that subjects can assume and the recruits are function of the recruiter characteristics.

- **Graphical structure** [Crawford, 2016]

A hidden population is an undirected graph, and we observe it partially in the *recruitment graph*, as also the coupon matrix and recruitment times. The unobserved graph is treated as *missing data* what can be interpreted as as Exponential Random Graph Model.

Markov chain model

- In a survey, questions about ethnicity, location, and gender create states in which each participant will be. By statistical tests, one can verify association between the recruiter and recruited responses.
- Heckathorn concluded that the recruitment was a memoryless process, and concluded RDS was a first-oder Markov process. At any instant, the current location of the point indicates the most recent recruit's characteristic.
- The markov chain must be ergodic.

Example of RDS in Markov chain

Consequences of Markov chain theory

Theorem

An equilibrium mix of recruits will be attained when the number of waves goes to infinity, and it is independent from which recruitment began.

Theorem

The subject pool generated approaches the equilibrium in a geometric rate.

Theorem (Remark)

If there is a substancial in-group selection bias, this convergence will take longer.

Assessing bias in RDS

Network model

- Let $G = (V, E)$ be an undirected graph representing the hidden population. The *Recruitment Graph* is $G_R = (V_R, E_R)$, where V_S represents the recruited individuals, and E_R the recruitment edge. The *Recruitment-induced Subgraph* is the induced subgraph by V_R.
- The *Coupon Matrix* C has elements $C_{ij} = 1$ if the subject i has at least one coupon just before the jth recruitment event.
- We observe $Y = (G_R, d, t, C)$.
- The time to recruitment along a *susceptible edge* has Exponential distribution, independent of the identity, neighbor, and all the other waiting times.

Example of observed data

Consequences

Theorem (Waiting time for a recruitment)

Let u be a recruiter and $v \in S_u$ a susceptible neighbor. The waiting time to u recruit v conditioned on the recruitment event has distribution Exponential with rate $\lambda|S_u|$. The probability of $v \in S_u$ to be the next recruited is uniform.

Theorem (Waiting time for some recruitment to occur)

The waiting time to the next recruitment is distributed as Exponential with rate $\lambda \sum_{u \in R} |S_u|$.

Likelihood of the recruitment time series

Table of Contents

Table of Contents

References I

Crawford, F. W. (2016).
The graphical structure of respondent-driven sampling.
Sociological Methodology, 46(1):187–211.

Deaux, E. and Callaghan, J. W. (1985).
Key informant versus self-report estimates of health-risk behavior.
Evaluation Review, 9(3):365–368.

Goodman, L. A. (1961).
Snowball Sampling.
The Annals of Mathematical Statistics, 32(1):148–170.

Heckathorn, D. D. (1997).
Respondent-driven sampling: A new approach to the study of hidden populations.
Social Problems, 44(2):174–199.

References II

Watters, J. K. and Biernacki, P. (1989).
Targeted sampling: Options for the study of hidden populations.
Social Problems, 36(4):416–430.