Chapter 1

Respondent-driven sampling

Respondent-driven sampling is a technique for estimating traits in hidden population. It is widely used for studying prevalence of HIV/AIDS among injection drug users, sex workers, men who have sex with men.

Studying prevalence of disease can help to understand and control its spreading. Unfortunately, there are difficulties with such kind of research as there is no sampling frame and members of hidden groups may not want reveal themselves.

There are several existing solutions for sampling hidden population such as snowball sampling, targeted sampling, time-space sampling, key-informant sampling. The main disadvantage of all these methods is unknown bias and variance of obtained estimation.

RDS begins with selecting group of initial participants that are called seeds. The procedure follows according to chain-referral model: each participant in study recruits another participants. The step is called wave. Both participating and recruiting new participants are encouraged by financial incentive. The sampling continues in this way until needed size of participants is reached. During RDS participants are asked to report how many contacts they have. This process enables to collect data for making statistical analysis.

In order to study formally RDS can be regarded as Markov Chain. Assumptions:

1. Seeds are chosen proportionally to their degree in the network.

2. If individual A knows individual B than individual B knows A as well (network can be represented as undirected graph).

3. The same individual can be recruited multiple times (sampling with replacement.

4. The choice of contacts to recruit is uniformly at random.

5. Individuals know precisely their network degree.

6. Each individual is reachable from each other individual (network is connected).

For this process stationary distribution is exactly distribution proportional to network degree. So first assumption guaranties that not only first but all samples during the process are taken with probability proportional to the degree of participants in the network. In [4] this assumption is considered to be reasonable as the people that are drawn as seeds are well-known people and they have usually more contacts than on average. Without this assumption first there should be performed enough number of waves until sample can be considered drawn from stationary distribution. simulation studies about assumptions violation(sensitivity) [1]

studies of variance

In this way individuals with more friends (contacts) are more likely to be recruited. To correct this bias the responses from individuals are weighted according to their degree (number of contacts). Let $X_1, X_2, ..., X_n$ be all collected samples during RDS. Then estimate μ_f of the population mean of f is defined [3] as

$$\mu_f = \frac{1}{\sum_{i=1}^{n} 1/degree(X_i)} \sum_{i=1}^{n} \frac{f(X_i)}{degree(X_i)}$$

RDS can perform poorly if the groups of individuals form different communities. It is known fact that friends tend to have similar traits. This fact becomes a source of bias in chain-referral methods of sampling. Structure of network also affects a lot. In [2] it is shown that 'bottlenecks' between different groups in hidden population increases variance of RDS estimator. They try RDS on network structure with communities, but where individuals, that are in contact with each other, do not have similar traits and showed that such structure indeed affects on RDS estimate.

Design effect d is variance of RDS estimate over variance of estimate obtained from simple random sampling (SRS). It means that if for SDS we

need n samples than to have RDS estimate with the same variance we need dn samples.

It is known fact that people tend to be friends if they share some traits: have similar age, common language, the same university.

Homophily - the tendency for individuals with similar attributes to be friends with one another. The fact that the majority of participants are recruited by other respondents and not by researchers makes RDS a successful method of data collection. However, the same feature also inherently complicates inference because it requires researchers to make assumptions about the recruitment process and the structure of the social network connecting the study population.

Bibliography

[1] Krista J Gile and Mark S Handcock. Respondent-driven sampling: An assessment of current methodology. *Sociological methodology*, 40(1):285–327, 2010.

[2] Sharad Goel and Matthew J Salganik. Respondent-driven sampling as markov chain monte carlo. *Statistics in medicine*, 28(17):2202–2229, 2009.

[3] Sharad Goel and Matthew J Salganik. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences*, 107(15):6743–6747, 2010.

[4] Matthew J Salganik and Douglas D Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology*, 34(1):193–240, 2004.