Karolinska Institutet

Department of Public Health Sciences

Respondent-Driven Sampling

Theory, Limitations & Improvements

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i MTC Lecture Hall on Theorells väg 1

Friday 22 Febuary, 2013, 09:00am

av

Xin Lu

Huvudhandledare:
Professor Fredrik Liljeros
Stockholm University
Department of Sociology

Bihandledare:
Docent Anna Ekéus Thorson
Karolinska Institutet
Department of Public Health Sciences

Doctor Monica Klinovoj Nordvik
Mid Sweden University
Department of Social Work

Fakultetsopponent:
Associate Professor Jari Saramäki
Aalto University
Department of Biomedical Engineering and Computational Science

Betygsnämnd:
Professor Karl Ekdahl
European Centre for Disease Prevention and Control (ECDC)

Doctor Anders Tegnell
Swedish Institute for Communicable Disease Control
Department of Analysis and Prevention

Docent Jette Möller
Karolinska Institutet
Department of Public Health Sciences

Stockholm 2013

Abstract

Background: The key purpose of sampling is to gain knowledge about a population using a small, affordable subset of selected individuals. This goal is often approached by choosing a representative sample with each individual's selection probability determined by a full list of individuals from the target population. However, for many populations central to the public health sciences, such as men who have sex with men (MSM), injecting drug users (IDUs), etc., the selection probability of individuals cannot be determined ahead of time because the list of all individuals is not available, impairing the generalization of results from the sample to the population. Respondent-driven sampling (RDS) was developed to generate representative samples of such hard-to-reach populations with improved accessibility. It provides an automated self-growing sampling design as well as asymptotically unbiased population estimates, making it the state-of-the-art sampling method for studying HIV-related key populations at risk in the past years. However, the availability of RDS estimates relies on many assumptions that are often not satisfied in real practice.

Aims: To assess the effect of violating assumptions on the performance of RDS estimators and to improve both the implementation and methodology of RDS for hard-to-reach populations of relevance to the public health sciences.

Contributions: The performance of RDS estimators is evaluated under various conditions. Results indicate that long chains initiated by diverse seeds are highly beneficial, while estimate bias is large if the network is directed or if respondents' participation behavior (such as preferential recruitment) depends on characteristics that are correlated with study outcomes. An Internet-based RDS (WebRDS) recruiting system is developed to circumvent the limitation of physical interview-based implementations. The system shows its ability to recruit sustaining location-free respondents in a study of MSM in Vietnam. Statistical methods are developed to generalize the RDS method from undirected networks to directed networks. The new method can function as a sensitivity test tool to account for the uncertainties of network directedness and error in self-reported degree data. Lastly, by integrating traditional RDS chain data with self-reported ego network data, a new estimator was developed to improve the reliability and validity of RDS. The new estimator shows not only improved precision, but also strong robustness to the preference of peer recruitment and variations in network structural properties.

Conclusions: Violations of assumptions are inevitable and should be investigated thoroughly in RDS practice. Due to the relatively high variance and vulnerability to certain harmful conditions, such as directedness, preferential recruitment, etc., results from RDS studies should be interpreted with caution. Researchers are encouraged to collect ego network data through the implementation of RDS to improve the precision of population estimates. In spite of its limited ability to generate close-enough population estimates, RDS is easily implementable and it offers a method with an improved response rate, providing an alternative to gain access/venue to the understanding of hard-to-access population.

Keywords: social networks, directed networks, ego networks, sampling, nonprobability sampling, respondent-driven sampling, Internet, estimator, bias, variance, public health, HIV, hidden population, differential recruitment, reporting error