15TH NATIONAL CONVENTION ON STATISTICS

03-05 OCTOBER 2022

Organized by the Philippine Statistical System
Spearheaded by the Philippine Statistics Authority

Estimation for Networked Hard-to-Reach Populations Under Respondent-Driven Sampling

Xavier Javines Bilon

School of Statistics

University of the Philippines Diliman

Computational Statistics

Crowne Plaza Galleria Manila

10:30–12:00, October 4, 2022

Hard-to-reach populations present a methodological challenge in obtaining samples and in drawing inferences about population characteristics from the samples obtained.

Studying hard-to-reach populations

Two characteristics distinguish hard-to-reach populations from other populations.

1. **undefined size and boundaries**
   → lack of an appropriate sampling frame or the impracticability of constructing one
      → probability sampling difficult or impossible to employ

2. membership in the population involves **stigmatized or illegal behavior**
   → individuals may refuse to cooperate or give unreliable answers to protect their privacy

Respondent-driven sampling

Heckathorn (1997) developed an alternative sampling method that utilizes networks connecting the members of hard-to-reach populations.

Respondent-driven sampling (RDS) is a variant of chain-referral sampling in which both sampling and estimation is facilitated by viewing the sample as a subset of the social network under study.

Respondent-driven sampling

RDS has gained popularity in recent years in studies involving hard-to-reach populations.

- people living with HIV (e.g., Malekinejad et al., 2008),
- gay men and people who inject drugs (e.g., Ramirez-Valles et al., 2005), and
- transgender persons (e.g., Bauer et al., 2015)

Respondent-driven sampling

Current RDS estimators rely on several strong assumptions in order to treat the sample as a probability sample.

- small sampling fraction
- accurate reporting of degree
- seed selection procedure and random selection of peers to recruit
- reciprocal relationships between peers

Dependence on these assumptions presents possible issues such as poor performance when they are applied to actual data and some of these assumptions are not met.

Volz & Heckathorn, 2008; Gile & Handcock, 2010; Goel & Salganik, 2010

Objective

In this study, we demonstrated the application of bootstrap-based procedures in estimating population characteristics of hard-to-reach populations using respondent-driven samples.

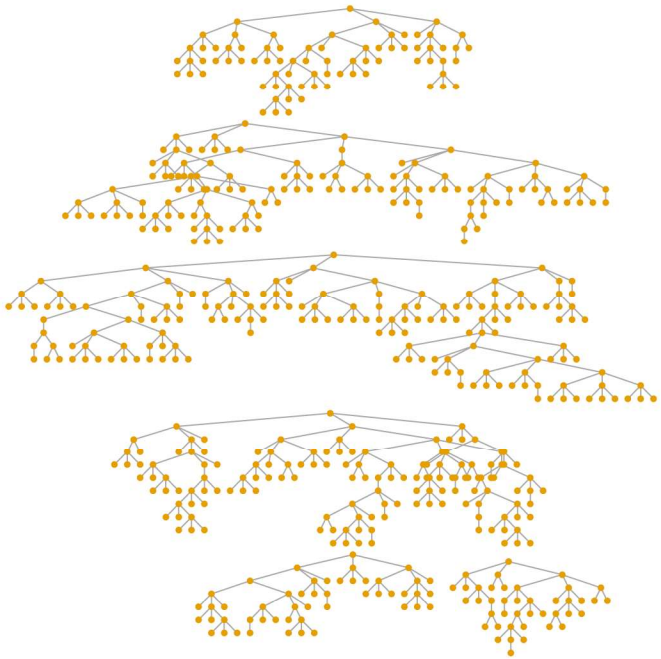
- Nonparametric bootstrap for homogeneous means (e.g., proportion and mean)
- Model-based and residual-based bootstrap for heterogeneous means (e.g., coefficients of linear, logistic, and Poisson regression models)

Overview of data sets

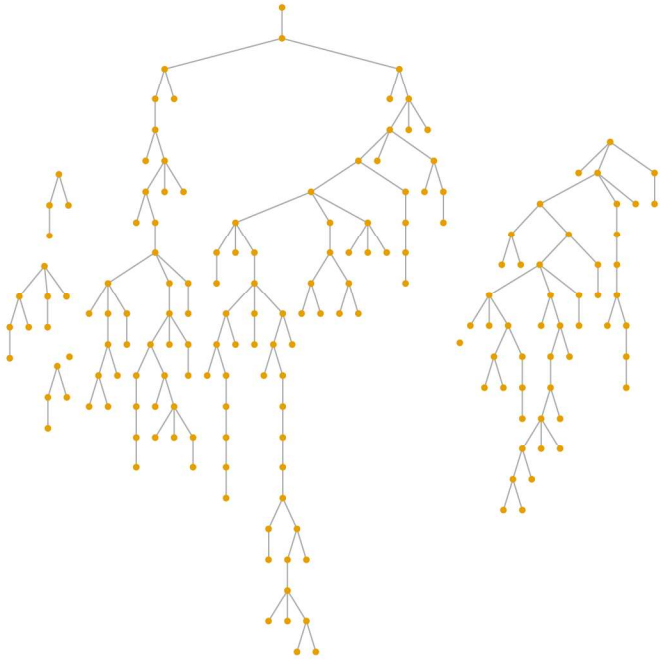
To demonstrate the application of the proposed methods in estimating parameters of networked hard-to-reach populations using respondent-driven samples, two data sets were explored in this study.

- **600 observations from people who inject drugs in a region in Estonia** (PWID Estonia data set; Wu et al., 2017)
- **176 observations from Syrian activist-refugees in Jordan** (Refugee Syria data set; Khoury, 2020a; Khoury, 2020b).

Recruitment trees

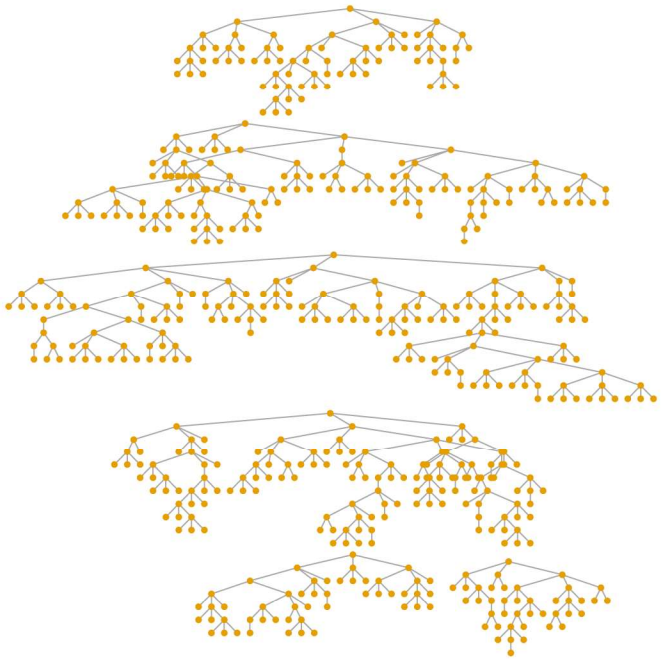


PWID Estonia Data Set

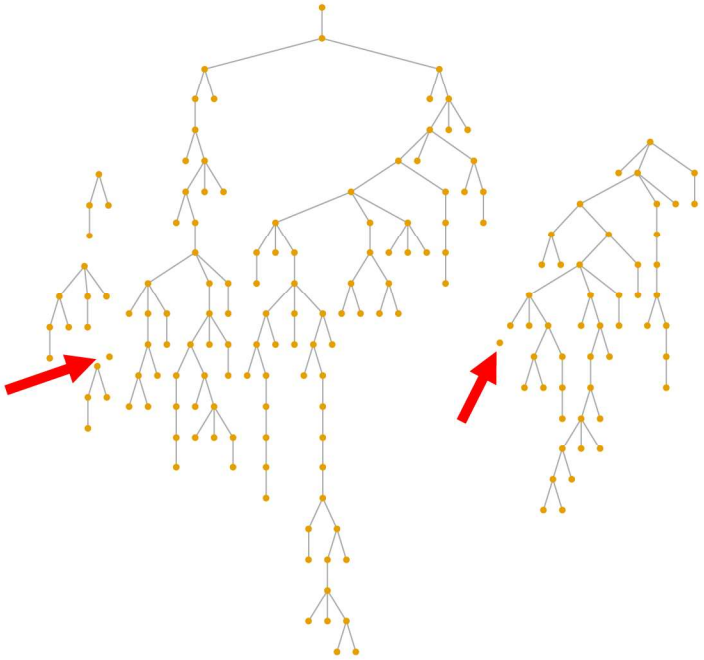


Refugee Syria Data Set

Recruitment trees



PWID Estonia Data Set



Refugee Syria Data Set

Overview of data sets

	PWID Estonia (Wu et al., 2017)	**Refugee Syria** (Khoury, 2020a; Khoury, 2020b)
Sample size	600	176
Number of seeds	6	7
Seed selection	Not mentioned	Purposive
Number of recruits	3	3
Number of degrees	Mdn = 15; Min = 2; Max = 100	Mdn = 12; Min = 1; Max = 75
Number of non-technical variables	1	21
Parameters to be estimated	proportion of people who inject drugs who were enrolled to ART in a region in Estonia	proportion of Syrian activist-refugees in Jordan who cooperated with other Syrian activists
		Logistic regression coefficients (whether a Syrian activist-refugee cooperated with other Syrian activists)

Overview of data sets

Outcome Variable		Frequency	Percentage
Whether a person who injected drugs was enrolled in ART in a region in Estonia (PWID Estonia Data Set)	Total	600	100.00%
	Yes	123	20.50%
	No	477	79.50%
Whether a Syrian activist-refugee in Jordan cooperated with other Syrian activists (Refugee Syria Data Set)	Total	171	100.00%
	Yes	156	91.23%
	No	15	8.77%

Overview of data sets

Predictor Variable		Frequency (Row Percentage)		
		Whether a Syrian activist-refugee in Jordan cooperated with other Syrian activists		
		Yes	No	Total
Assigned sex	Female	6 (11.11%)	48 (88.89%)	54 (100.00%)
	Male	9 (7.69%)	108 (92.31%)	117 (100.00%)
Participation in activism in Syria before 2011	Yes	13 (9.70%)	121 (90.30%)	134 (100.00%)
	No	2 (5.41%)	35 (94.59%)	37 (100.00%)
Participation in activism in Syria after 2011 but before coming to Jordan	Yes	9 (10.00%)	81 (90.00%)	90 (100.00%)
	No	6 (7.41%)	75 (92.59%)	81 (100.00%)
Agree that most people can be trusted	Yes	14 (9.27%)	137 (90.73%)	151 (100.00%)
	No	1 (5.00%)	19 (95.00%)	20 (100.00%)
Predictor Variable		**Descriptive Statistics**		
Age	Mean	27.74		
	Median	26.00		
	SD	6.72		

Point estimates and estimated standard errors

Parameter	RDS II	Nonparametric Bootstrap
Point Estimate		
Proportion (PWID Estonia)	16.01%	20.48%
Proportion (Refugee Syria)	82.26%	91.48%
Estimated Standard Error		
Proportion (PWID Estonia)	1.87%	1.71%
Proportion (Refugee Syria)	5.64%	2.15%

Point estimates and estimated standard errors

Parameter	LSE (RDS)	Model-based Bootstrap	Residual-based Bootstrap
Point Estimate			
Intercept	2.2927	2.3722	2.2684
Assigned sex (Male)	0.3652	0.4005	0.3872
Age	−0.0170	−0.0144	−0.0177
Participation in activism in Syria before 2011 (Yes)	0.6749	2.6415	0.7215
Participation in activism in Syria after 2011 but before coming to Jordan (Yes)	0.2693	0.3218	0.2869
Agree that most people can be trusted (Yes)	0.5519	5.4397	0.5980
Estimated Standard Error			
Intercept	1.2414	1.7743	0.2974
Assigned sex (Male)	0.5731	0.7011	0.0800
Age	0.0391	0.0508	0.0051
Participation in activism in Syria before 2011 (Yes)	0.7919	5.4621	0.1119
Participation in activism in Syria after 2011 but before coming to Jordan (Yes)	0.5634	1.3446	0.0710
Agree that most people can be trusted (Yes)	1.0745	7.6923	0.1601

Point estimates and estimated standard errors

Parameter	LSE (RDS)	Model-based Bootstrap	Residual-based Bootstrap
Point Estimate			
Intercept	2.2927	2.3722	2.2684
Assigned sex (Male)	0.3652	0.4005	0.3872
Age	−0.0170	−0.0144	−0.0177
Participation in activism in Syria before 2011 (Yes)	0.6749	2.6415	0.7215
Participation in activism in Syria after 2011 but before coming to Jordan (Yes)	0.2693	0.3218	0.2869
Agree that most people can be trusted (Yes)	0.5519	5.4397	0.5980
Estimated Standard Error			
Intercept	1.2414	1.7743	0.2974
Assigned sex (Male)	0.5731	0.7011	0.0800
Age	0.0391	0.0508	0.0051
Participation in activism in Syria before 2011 (Yes)	0.7919	5.4621	0.1119
Participation in activism in Syria after 2011 but before coming to Jordan (Yes)	0.5634	1.3446	0.0710
Agree that most people can be trusted (Yes)	1.0745	7.6923	0.1601

Point estimates and estimated standard errors

Parameter	LSE (RDS)	Model-based Bootstrap	Residual-based Bootstrap
Point Estimate			
Intercept	2.2927	2.3722	2.2684
Assigned sex (Male)	0.3652	0.4005	0.3872
Age	−0.0170	−0.0144	−0.0177
Participation in activism in Syria before 2011 (Yes)	0.6749	2.6415	0.7215
Participation in activism in Syria after 2011 but before coming to Jordan (Yes)	0.2693	0.3218	0.2869
Agree that most people can be trusted (Yes)	0.5519	5.4397	0.5980
Estimated Standard Error			
Intercept	1.2414	1.7743	0.2974
Assigned sex (Male)	0.5731	0.7011	0.0800
Age	0.0391	0.0508	0.0051
Participation in activism in Syria before 2011 (Yes)	0.7919	5.4621	0.1119
Participation in activism in Syria after 2011 but before coming to Jordan (Yes)	0.5634	1.3446	0.0710
Agree that most people can be trusted (Yes)	1.0745	7.6923	0.1601

Point estimates and estimated standard errors

Parameter	LSE (RDS)	Model-based Bootstrap	Residual-based Bootstrap
Point Estimate			
Intercept	2.2927	2.3722	2.2684
Assigned sex (Male)	0.3652	0.4005	0.3872
Age	−0.0170	−0.0144	−0.0177
Participation in activism in Syria before 2011 (Yes)	0.6749	2.6415	0.7215
Participation in activism in Syria after 2011 but before coming to Jordan (Yes)	**0.2693**	**0.3218**	**0.2869**
Agree that most people can be trusted (Yes)	0.5519	5.4397	0.5980
Estimated Standard Error			
Intercept	1.2414	1.7743	0.2974
Assigned sex (Male)	0.5731	0.7011	0.0800
Age	0.0391	0.0508	0.0051
Participation in activism in Syria before 2011 (Yes)	0.7919	5.4621	0.1119
Participation in activism in Syria after 2011 but before coming to Jordan (Yes)	**0.5634**	**1.3446**	**0.0710**
Agree that most people can be trusted (Yes)	1.0745	7.6923	0.1601

Point estimates and estimated standard errors

Parameter	LSE (RDS)	Model-based Bootstrap	Residual-based Bootstrap
Point Estimate			
Intercept	2.2927	2.3722	2.2684
Assigned sex (Male)	0.3652	0.4005	0.3872
Age	−0.0170	−0.0144	−0.0177
Participation in activism in Syria before 2011 (Yes)	0.6749	2.6415	0.7215
Participation in activism in Syria after 2011 but before coming to Jordan (Yes)	0.2693	0.3218	0.2869
Agree that most people can be trusted (Yes)	**0.5519**	**5.4397**	**0.5980**
Estimated Standard Error			
Intercept	1.2414	1.7743	0.2974
Assigned sex (Male)	0.5731	0.7011	0.0800
Age	0.0391	0.0508	0.0051
Participation in activism in Syria before 2011 (Yes)	0.7919	5.4621	0.1119
Participation in activism in Syria after 2011 but before coming to Jordan (Yes)	0.5634	1.3446	0.0710
Agree that most people can be trusted (Yes)	**1.0745**	**7.6923**	**0.1601**

Conclusions

- Illustrative results suggest that bootstrap-based procedures may perform better than existing RDS estimation procedures in making inferences about population characteristics of hard-to-reach populations under respondent-driven sampling.

- Simulation studies to further demonstrate the applicability of bootstrap-based procedures in estimating parameters of hard-to-reach populations using respondent-driven samples

References

Bauer, G. R., Scheim, A. I., Pyne, J., Travers, R., & Hammond, R. (2015). Intervenable factors associated with suicide risk in transgender persons: A respondent driven sampling study in Ontario, Canada. BMC Public Health, 15(1), 525. https://doi.org/10.1186/s12889-015-1867-2

Beaudry, I. S., Gile, K. J., & Mehta, S. H. (2017). Inference for respondent-driven sampling with misclassification. The Annals of Applied Statistics, 11(4), 2111-2141. https://doi.org/10.1214/17-AOAS1063

Braunstein, M. S. (1993). Sampling a hidden population: Noninstitutionalized drug users. AIDS Education and Prevention: Official Publication of the International Society for AIDS Education, 5(2), 131-140.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970319

Gile, K. J., & Handcock, M. S. (2010). 7. Respondent-driven sampling: An assessment of current methodology. Sociological Methodology, 40(1), 285-327. https://doi.org/10.1111/j.1467-9531.2010.01223.x

Goel, S., & Salganik, M. J. (2009). Respondent-driven sampling as Markov chain Monte Carlo. Statistics in Medicine, 28(17), 2202-2229. https://doi.org/10.1002/sim.3613

Goel, S., & Salganik, M. J. (2010). Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences, 107(15), 6743-6747. https://doi.org/10.1073/pnas.1000261107

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems, 44(2), 174-199. https://doi.org/10.2307/3096941

Heckathorn, D. D. (2002). Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Social problems, 49(1), 11-34. https://doi.org/10.1525/sp.2002.49.1.11

Heckathorn, D. D., Broadhead, R. S., & Sergeyev, B. (2001). A methodology for reducing respondent duplication and impersonation in samples of hidden populations. Journal of Drug Issues, 31(2), 543-564. https://doi.org/10.1177/002204260103100209

Kanouse, D. E., Berry, S. H., Duan, N., Lever, J., Carson, S., Perlman, J. F., & Levitan, B. (1999). Drawing a probability sample of female street prostitutes in Los Angeles County. Journal of Sex Research, 36(1), 45-51. https://doi.org/10.1080/00224499909551966

Kerr, L., Kendall, C., Guimarães, M. D. C., Mota, R. S., Veras, M. A., Dourado, I., ... & Johnston, L. G. (2018). HIV prevalence among men who have sex with men in Brazil: results of the 2nd national survey using respondent-driven sampling. Medicine, 97(1 Suppl). https://doi.org/10.1097/MD.0000000000010573

Khoury, R. B. (2020a). Hard-to-survey populations and respondent-driven sampling: Expanding the political science toolbox. Perspectives on Politics, 18(2), 509-526. https://doi.org/10.1017/S1537592719003864

Khoury, R. B. (2020b). Replication data for: Hard-to-survey populations and respondent-driven sampling: Expanding the political science toolbox (Version 1) [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/XKOVUN

Malekinejad, M., Johnston, L. G., Kendall, C., Kerr, L. R. F. S., Rifkin, M. R., & Rutherford, G. W. (2008). Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. AIDS and Behavior, 12(1), 105-130. https://doi.org/10.1007/s10461-008-9421-1

Ramirez-Valles, J., Heckathorn, D. D., Vázquez, R., Diaz, R. M., & Campbell, R. T. (2005). From networks to populations: The development and application of respondent-driven sampling among IDUs and Latino gay men. AIDS and Behavior, 9(4), 387-402. https://doi.org/10.1007/s10461-005-9012-3

Salganik, M. J. (2006). Variance estimation, design effects, and sample size calculations for respondent-driven sampling. Journal of Urban Health: Bulletin of the New York Academy of Medicine, 83(7), i98-i112. https://doi.org/10.1007/s11524-006-9106-x

Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. Sociological Methodology, 34(1), 193-240. https://doi.org/10.1111/j.0081-1750.2004.00152.x

Volz, E., & Heckathorn, D. D. (2008). Probability based estimation theory for respondent driven sampling. Journal of Official Statistics, 24(1), 79-97.

Wu, J., Crawford, F. W., Raag, M., Heimer, R., & Uusküla, A. (2017). Using data from respondent-driven sampling studies to estimate the number of people who inject drugs: Application to the Kohtla-Järve region of Estonia. PLOS one, 12(11). https://doi.org/10.1371/journal.pone.0185711

15TH NATIONAL CONVENTION ON STATISTICS

03-05 OCTOBER 2022

Organized by the Philippine Statistical System
Spearheaded by the Philippine Statistics Authority

Estimation for Networked Hard-to-Reach Populations Under Respondent-Driven Sampling

Xavier Javines Bilon

School of Statistics

University of the Philippines Diliman

E-mail: xjbilon@up.edu.ph

Twitter: @xjbilon

Website: sites.google.com/up.edu.ph/xjbilon

Thank you!

http://www.psa.gov.ph/ncs

http://openstat.psa.gov.ph

https://twitter.com/PSAgovph

https://www.facebook.com/PhilippineStatisticsAuthority