Testing the Random Recruitment Assumption of Respondent-Driven Sampling: Practical Implications

Jichuan Wang, Ph.D., Linna Li, M.S
Russel S. Falck, M.A., Robert G. Carlson, Ph.D.

Center for Intervention, Treatment and Addiction Research
Boonshoft School of Medicine
Wright State University

Paper presented in AHPA 135th Annual Meeting November 3-7, 2007, Washington, D.C.

Abstract

Respondent-driven sampling (RDS) has been increasingly applied to sample hidden populations, such as illicit drug users. In addition to a number of other advantages, RDS sample analysis provides asymptotically unbiased population composition estimation, which can be used to weight the sample for producing unbiased sample statistics, as well as for population size estimation. In the current practice of RDS sample analysis, sample recruitment patterns are used as the estimates of personal network compositions that are used for population proportions estimation. The precision of replacing network compositions with sample recruitment patterns relies on the assumption of random recruitment from personal networks. Although this assumption serves as a basis of RDS analysis, testing of this assumption has been rare. A SAS macro has been developed by the authors to conduct such a test in RDS sample analysis, using bootstrap method. Real data are used for demonstration.

Population Composition Estimation in RDS Sample Analysis

• In RDS sample analysis, asymptotically unbiased population compositions of a hidden population can be estimated for the purpose of producing unbiased sample statistics.

• Population compositions are treated as a function of two components: personal network compositions and mean "degree."

$$P_A = \frac{D_B \cdot C_{BA}}{D_A \cdot C_{AB} + D_B \cdot C_{BA}}$$

$$P_B = \frac{D_A \cdot C_{AB}}{D_A \cdot C_{AB} + D_B \cdot C_{BA}} \qquad (1)$$

• Estimation of average size of personal networks:

$$\hat{D}_A = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}} \qquad (2)$$

• Estimation of personal network compositions:

$$S_{AB} = \frac{r_{AB}}{r_{AA} + r_{AB}}$$

$$S_{BA} = \frac{r_{BA}}{r_{BB} + r_{BA}} \qquad (3)$$

• The formulas actually used for population proportion estimation in RDS sample analysis:

$$\hat{P}_A = \frac{\hat{D}_B \cdot S_{BA}}{\hat{D}_A \cdot S_{AB} + \hat{D}_B \cdot S_{BA}}$$

$$\hat{P}_B = \frac{\hat{D}_A \cdot S_{AB}}{\hat{D}_A \cdot S_{AB} + \hat{D}_B \cdot S_{BA}} \qquad (4)$$

- Personal network compositions are estimated from sampling recruitment patterns that represent the links in personal networks.
- It is assumed that respondents recruit their peers randomly from their personal networks.
- Although, this fundamental assumption serves as a basis for population proportion estimation in RDS sample analysis, testing this assumption is rare.

• Non-random recruitment often occurs in chain-referral sampling because of the effects of masking and volunteerism.

• RDS is a modified version of chain-referral sampling method, which is designed to reduce the masking and volunteerism effects by employing a dual incentive system and referral coupons to guide peer to peer recruitment.

• However, the effects of masking and volunteerism may not be entirely excluded, although they may be substantially reduced,

• Other factors, such as social and geographic proximities, may also influence the way respondents recruit from their personal networks.

• It is necessary for RDS practitioners to test the assumption of random recruitment in their RDS sample analysis.

Testing the Assumption of Random Recruitment

• The assumption of random recruitment can be statistically tested by comparing the sampling recruitment patterns with the self-reported network compositions of trait groups.

• t-test (Wang et al., 2005, 2007).

• bootstrap method to estimate s.e. of discrepancy between recruitment patterns and self-reported personal network compositions:

$$\hat{\sigma}_{\hat{\theta}} = \left\{ \frac{\sum \left(\hat{\theta}_b - \hat{\theta}_{(.)} \right)^2}{(B-1)} \right\}^{1/2} \qquad (5)$$

• A SAS macro has been developed by the authors for RDS sample analysis and testing the assumption of random recruitment from personal networks.

Examples

• **Sample 1**: MDMA/ecstasy users (N=402) (May 2002 and June 2003).

- The results of comparisons between recruitment patterns and respondent reported personal network compositions are shown in Table1 1. The statistical tests were based on 1000 bootstrap resamples.

Table 2. Comparisons between sample recruitment patterns and personal network compositions among MDMA users (n=343)[1]

Recruiter		Recruits		
Gender		Male	Female	Total
Male (n=211)	Recruitment[2]	161 (67.0%)	79 (33.0%)	240 (100.0%)
	Network[3]	4,349 (54.3%)	3,661 (45.7%)	8,010 (100.0%)
	Difference	12.7%	-12.7%	
	T-test[4]	**t = 3.8097 p = 0.0001**	**t = -3.8097 p = 0.0001**	
Female (n=132)	Recruitment[2]	50 (48.4%)	53 (47.6%)	103 (100.0%)
	Network[3]	1,927 (52.2%)	1,765 (47.8%)	3,692 (100.0%)
	Difference	-3.7%	3.7%	
	T-test[4]	t = -0.7378 p = 0.4608	t = 0. 7378 p = 0.4608	
Ethnicity		White	Non White	Total
White (n=272)	Recruitment[2]	244 (85.2%)	42 (14.8%)	286 (100.0%)
	Network[3]	8,249 (87.4%)	1,189 (12.6%)	9,438 (100.0%)
	Difference	-2.2%	2.2%	
	T-test[4]	t = -0.8580 p =0.3911	t = 0. 8580 p =0.3911	
Non White (n=71)	Recruitment[2]	28 (49.2%)	29 (50.8%)	57 (100.0%)
	Network[3]	1,579 (68.6%)	723 (31.4%)	2,302 (100.0%)
	Difference	-19.4%	19.4%	
	T-test[4]	**t =-2.4828 p = 0.0132**	**t = 2. 4828 p = 0.0132**	

Notes.

[1] - RDS sample of MDMA users recruited from the Columbus area in Ohio during May 2002 and June 2003. Among the total sample of 402 MDMA users, 28 were seeds who were used to initialize the sampling process. Because seeds were selected with a different mechanism, they were excluded from RDS sample analysis. Since data collection on the personal networks did not begin until several weeks after the sampling process was initiated, only 343 participants were available for this study.

[2] - Sample recruitment pattern.

[3] - Self-reported personal network size by group.

[4] - Bootstrap standard error based on 1000 resamples was used to test the difference between recruitment patterns and personal network compositions.

• **Sample 2**: Rural stimulant users (N=248)(October 2002 and March 2004).

- The results of comparisons between recruitment patterns and respondent reported personal network compositions are shown in Table1 2. The statistical tests were based on 1000 bootstrap resamples.

Table 3. Comparisons between sample recruitment probability and personal network compositions among rural stimulant users (n=230)[1]

Recruiter		Recruits		
Gender		Male	Female	Total
Male (n=155)	Recruitment[2]	101 (71.6%)	40 (28.4%)	141 (100.0%)
	Network[3]	4,526 (61.0%)	2,894 (39.0%)	7,420 (100.0%)
	Difference	10.6%	-10.6%	
	T-test	**t = 2.4765 p = 0.0134**	**t = -2.4765 p = 0.0134**	
Female (n=75)	Recruitment[2]	54 (60.6%)	35 (39.4%)	89 (100.0%)
	Network[3]	1,398 (57.6%)	1,029 (42.4%)	2,427 (100.0%)
	Difference	3.0%	-3.0%	
	T-test	t = 0.5281 p = 0.5975	t =-0.5281 p = 0.5975	
Ethnicity		White	Non White	Total
White (n=203)	Recruitment[2]	191 (92.8%)	15 (7.2%)	206 (100.0%)
	Network[3]	6,302 (73.6%)	2,260 (26.4%)	8,562 (100.0%)
	Difference	19.2%	-19.2%	
	T-test	**t = 5.4314 p <0.0001**	**t = -5.4314 p <0.0001**	
Non White (n=27)	Recruitment[2]	12 (49.4%)	12 (50.6%)	24 (100.0%)
	Network[3]	659 (54.3%)	554 (45.7%)	1,213 (100.0%)
	Difference	-4.9%	4.9%	
	T-test	t =-0.3711 p = 0.7106	t = 0.3711 p = 0.7106	

Notes.
[1] - RDS sample of rural stimulant users recruited from three contiguous rural counties in west-central Ohio between October 2002 and March 2004. Among the total sample of 248 stimulant users, 19 seeds excluded from RDS sample analysis.
[2] - Sample recruitment pattern.
[3] - Self-reported personal network size by group.
[4] - Bootstrap standard error based on 1000 resamples was used to test the difference between recruitment patterns and personal network compositions.

Conclusion

• The assumption of random recruitment in RDS sampling may hold for some trait groups, but may not for some other groups.

• Caution should be exercised in reporting and using the estimated population proportion of a trait group for the purpose of sample weighting and population size estimation when the assumption of random recruitment does not hold for the group.

• Even though the assumption of random recruitment does not hold for some respondent groups, a number of advantages of RDS remain:

1) no random seed selection is necessary for RDS; 2) sample compositions converge and reach equilibrium quickly independent of the characteristics of the initial sample or seeds; 3) RDS can reduce the effects of masking and volunteerism; 4) RDS sample analysis provides information about social structures in which members of the target population are embedded; 5) and finally, RDS is generally easier as well as less expensive to implement, compared with other sampling methods that employ full-time outreach workers.