Assessment and Validation of Respondent Driven Sampling

Jing Li,[a] M. Giovanna Merli,[a] Erik Nordheim[b] William Whipple Neely[b]

a) Department of Sociology and Center for Demography and Ecology, University of Wisconsin, Madison
b) Department of Statistics and Center for Demography and Ecology, University of Wisconsin, Madison

Introduction – Sampling female sex workers in Shanghai

We are engaged in a study of the social determinants of HIV/STDs in China which features an upcoming survey of sexual behavior and sexual networks among female sex workers (FSWs), urban residents and rural migrants in Shanghai. In this paper we discuss the challenges of sampling FSWs in Shanghai and evaluate the statistical validity of Respondent Driven Sampling (RDS), an approach that is becoming increasingly popular to sample hidden and hard-to-reach populations, such as female sex workers in Shanghai.

Most methods relating to the sampling of populations rely on "probability sampling" in which the probability of any possible sample can be quantified. For target groups whose members do not congregate in high proportions in identifiable or accessible locations, constructing a relatively complete sampling frame which would ensure complete coverage and an unbiased representation of the populations under study is not feasible. This is especially relevant for hard to reach or stigmatized populations, such as female sex workers (FSWs), injection drug users (IDUs) or gay populations in cultural settings where this behavior is highly stigmatized.

Generating a representative sample of FSWs in Shanghai presents challenges that are typical of research on hard-to-reach and stigmatized populations. In a city like Shanghai the difficulties appear even greater than in other locations with FSWs. For example, in the U.S. sex workers operate in geographically concentrated areas and make up a large proportion of women on the street (e.g. Berry, Duan and Kanouse 1990; Kanouse et al. 1999). In China, FSWs operate in karaoke bars, coffee shops, beauty and massage parlors, from the halls of hotels where they lure potential clients by calling their rooms directly, but also outside of public view as escorts accompanying clients for a fixed duration, as second wives of men with money and influential positions, by soliciting potential clients on the streets or in parks, in connivance with hotel workers, and by providing sexual services to the transient labor force of male workers living in urban slum dwellings on construction sites (Hershatter 1997; Pan 1997; Huang et al. 2004). This typology indicates that sex workers operate under different degrees of visibility and accessibility. Those who are establishment-based are, in principle, more accessible because they can be mapped through identifiable locations to provide a sampling frame. This is not the case for other types of FSWs who operate in less identifiable locations. In addition to these difficulties, prostitution in China is an illicit and highly stigmatized behavior. This makes FSWs reluctant to take part in formalized studies without having established a comfortable level of trust with the researchers.

The majority of epidemiological studies of risk behavior among FSWs in China have so far relied on convenience samples or on sampling frames established through the mapping of entertainment venues (Pan 1997; Huang et al. 2004) or through lists of FSWs in re-education camps (Xia and Yang 2005). Since the coverage of these frames is only

limited to FSWs who operate in public view, these studies underestimate risk behaviors because, at best, they allow inference to venue based sex workers who are less vulnerable as they have the option to refuse to provide sexual services or are in a better position to negotiate the use of condoms. On the other hand, FSWs who are not directly mediated through China's commercial recreational business sector and who sell sex by soliciting clients on the streets, in parks, or on construction sites are characterized by a more straightforward exchange of sex for financial or material recompense. A lone attempt to recruit more broadly from among all categories of FSWs was a study which employed snowball sampling (Ding et al. 2005) to recruit respondents. Beside venue based sex workers, this study was able to reach street workers, who operate in less identifiable locations. This study noted significant segmentation of sex work. Compared with venue based sex workers, street workers were older, less educated, more likely to be married, more likely to be of rural origin and reported more high risk behaviours, such as lower condom use, a higher mean number of sex partners per week and a history of STDs.

Respondent Driven Sampling to sample hard to reach populations

"Snowball sampling" (Goodman 1961), a "non-probability" approach, uses initial respondents to recruit other members to the sample. Hitherto, this approach has had mixed success (Cornelius 1982; Kalton and Anderson 1986). Recently, Respondent Driven Sampling (RDS) has been developed as a specific variant of snowball sampling (Heckathorn 1997, 2002). Like snowball sampling, it is predicated on the idea that referral made by members of the target population is more efficient than recruitment by researchers. RDS is based on certain assumptions about the social network connecting members of the population. In RDS, the sample is used to make estimates of the key parameters describing the social network. These estimates are combined with observed proportions to estimate the "true" proportions. RDS developers claim that inference from the sample to all potential members of the population is valid and that the method yields statistically unbiased estimates (Salganik and Heckathorn 2004). RDS has been used in studies of injection drug users (IDUs) in the U.S. and elsewhere, Latino gay populations in the U.S. (Ramirez Valles et al.), and FSWs in Vietnam (Johnston et al.1996, Chau et al.1996). The U.S. Center for Disease Control (CDC) has adopted RDS to track HIV-risk behaviors of injection drug users in 25 U.S. cities.

For sampling of hard-to-reach populations, we view the RDS approach as a useful step forward. However, RDS is not free from controversies surrounding it (Heimer 2005; Martin et al. 2003). But given its newness, a systematic assessment of this approach has yet to be undertaken.

Methods

There are many questions about RDS that have not been satisfactorily answered. A key issue concerns the validity of a number of its assumptions: e.g., (a) equal probability that a respondent, already contacted, will refer any of the individuals from her social network; (b) reciprocity (if individual X is linked to Y, then Y is linked to X); (c) accurate self-report of "degree" (i.e. number of friendships); (d) the network as one single 'connected' component. These assumptions require that social networks have the following characteristics: (1) Members of the population must know other like themselves; (2) networks must be dense to sustain a chain referral process, because otherwise recruitment

chain would die out quickly; (3) networks must not be biased towards the inclusions of individuals with interrelationship: they must avoid the situation where the most popular members of a population are more likely to be identified than others, while the more isolated members are missed; (4) networks must be fluid. For example, in highly segmented populations, cross-over between subgroups becomes impossible and recruitment remains trapped within some subgroups while there is no representation of other subgroups; (5) referrals must not be dependent on the subjective choices of the respondents first accessed.

Here, we evaluate the validity of these assumptions in the context of female sex workers (FSWs). We do this with a simulation model that is able to describe potential populations of FSWs that includes a realistic structure of their social networks. By modifying the model parameters, we evaluate what effects the possible violation(s) of the assumptions have on the inference that can be drawn. Variation in model parameters is empirically informed by results of exploratory research among female sex workers in Shanghai undertaken to explore barriers to network based sampling among this population.[1] However, we also explore a wider range of input parameters to allow our results to be generalizable to a broader range of populations.

The simulation model we have constructed allows us to evaluate what effect the violation of the assumptions has on the inference that can be drawn from an RDS study. The model is able to describe potential populations of FSWs that includes a realistic structure of their social networks. In such a model, different inputs can be considered and their effect on the statistical inference assessed. As an example, suppose that a key attribute of Shanghai FSWs about which we wish to learn is condom use. For a "known" population from which we simulate, we can vary input parameters, such as the probability of a condom-using FSW to recruit another condom-using FSW, to determine the impacts on the estimated proportion of condom users. By modifying parameters of the simulation model, we are able to assess the impact of the assumptions on inference.

The data generation portion of the simulation program is based on a Markovian process. This signifies that the probability of an event at time t+1 depends only on information from time t and not on prior information. Depending on the number of categories in which individuals are classified, two for condom use but potentially more for other possible outcome variables, the probability with which a recruiter from some

[1] We identified the following barriers to network based sampling of FSWs: (1) networks of FSWs may not fluid. Segmentation of sex work limits opportunities for social cross-over to other types, especially between FSWs who operate in identifiable locations and those who do not; (2) networks may not be dense: stigmatization associated with prostitution or FSWs' segregated lives may limit the network size of sex workers. Also, high levels of independence among certain types of sex workers may produce very small or inexistent networks; (3) networks of FSWs may be altered by the relationship with their pimps. Pimps are likely to interfere with the recruitment process if they see the project as a business opportunity and require the investigator to go through their own social networks of FSWs. This will limit the fluidity of sex workers' network and introduce the potential for preferential recruitment; (4) referrals will depend on the subjective perception of respondents about the involvement of others in a highly stigmatized activity.

category recruits a recruitee from the same or a different category is specified by a transition matrix. The probabilities in the matrix are altered to represent different structures of the social network. Another portion of the data generation determines the degree (number of friendships) for each individual in the study; this is modeled by a Poisson distribution with varying parameters. There are two other parameters that correspond to the specific conduct of the sampling scheme that will be varied.

From each generated data set, inference on quantities of inference (e.g., proportion of sex workers who use condoms) are drawn. Multiple data sets are generated from each set of conditions (transition probabilities, Poisson means, etc.) to assess the bias and variability of the inference. By assessing the bias and variance from a range of conditions, we determine the impact of the assumptions on the inference.

We will follow these simulation results with a field study to assess the validity of the assumptions. We anticipate that our work may serve as a template for those studying hard-to-reach populations in other settings with different social networks.

References

Berry, Sandra H., Naihua Duan and David E. Kanouse. 1990. "Developing a Probability Sample of Prostitutes: Sample Design for the RAND Study of HIV Infection and Risk Behaviors in Prostitutes." *Rand Note N-3190-NICHD*. Los Angeles, CA: Rand.

Cornelius, Wayne A. 1982 "Interviewing Undocumented Immigrants: Methodological Reflections Based on Fieldwork in Mexico and the United States," *International Migration Review,* 16(2):378-411.

Ding, Yanpeng et al. 2005. HIV Infection and Sexually Transmitted Diseases in Female Commercial Sex Workers in China. *Journal of Acquired Immune Deficiency Syndrome* 38(3):314-319.

Hershatter, Gail. 1997. *Dangerous Pleasures*. Berkeley, CA: University of California Press.

Huang, Yingying, Gail E. Henderson, Suiming Pan, and Myron S. Cohen. 2004. HIV/AIDS risk among brothel-based female sex workers in China: Assessing the terms, content and knowledge of sex work. *Sexually Transmitted Diseases* 31(11):695-700.

Goodman, Leo A. 1961. "Snowball sampling". *Annals of Mathematical Statistics* 32:148-70.

Heckathorn, Douglas D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44: 174-199.

Heckathorn, Douglas D. 2002. Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Social Problems* 49:11-34.

Heimer, Robert. 2005. Critical issues and further questions about respondent-driven sampling: Comment on Ramirez-Valles, et al. (2005). *AIDS and Behavior* 9(4): 403-408.

Johnston, Lisa G. et al. 2006. "Assessment of Respondent Driven Sampling for Recruiting Female Sex Workers in Two Vietnamese Cities: Reaching the Unseen Sex Worker." Unpublished manuscript.

Kalton, G. and D. Anderson. 1986. "Sampling Rare Populations." *Journal of the Royal Statistical Society, Series A* 149:65-82.

Kanouse, David E. et al. 1999. "Drawing a Probability Sample of Female Street Prostitutes in Los Angeles County." *The Journal of Sex Research* 36(1):45-51.

Martin, John Levi, James Wiley and Dennis Osmond. 2003. Social networks and unobserved heterogeneity in risks for AIDS. *Population Research and Policy Review* 22:65-90.

Pan, Suiming.1997. *Three Red Light Districts in South China*. (In Chinese). Guangzhou: Qunyan Chubandshe.

Ramirez-Valles, J., Heckathorn, D. D., Vázquez, R., Diaz, R. M., Campbell, R. T. 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

Salganik, Matthew J. and Douglas D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34:193-239.

Xia, Guomei and Xiushi Yang. "Risky Sexual Behavior among Female Entertainment Workers in China: Implications for HIV/STI Prevention Intervention," AIDS Education and Prevention, 17(2): 143-156, 2005.