NETWORK SAMPLING AS AN APPROACH TO SAMPLING PREGNANT WOMEN

Linda L. Sanders and William D. Kalsbeek, University of North Carolina
William D. Kalsbeek,Survey Research Unit, Department of Biostatistics,
University of North Carolina, Chapel Hill, NC 27599

KEY WORDS: Multiplicity sampling, Rare populations, Network sampling

1. INTRODUCTION

Network sampling is a strategy used in surveys to increase the discovery rate for persons with a relatively rare attribute. When effectively implemented it helps to reduce the sampling error for estimates of the proportion of the population with the trait and for estimates of other measures obtained for those with the trait. In principle it works by expanding the linkage of units on a sampling frame to additional members of the population, thus increasing the likelihood of discovering persons with the attribute of interest. In conventional surveys only respondents or members of their immediate household are screened for the trait, whereas in network designs the respondent becomes an informant for the household as well as some well-defined social network of which he or she is a part.

Somewhat ironically the origins of network sampling trace back to the development of theory aimed at combating the effects of the multiple frame linkage that it exploits. The early work by Sirken (1970, 1972a, 1972b) and then later by Nathan (1976), Levy (1977) and others paved the way for improved estimation in the presence of "multiplicity," as the problem came to be known. The irony was that while multiplicity presents an estimation issue with which one must contend, its existence on the frame, especially when contrived, actually benefits efforts to screen for rare population traits by enabling the investigator to "cast the net more broadly" from the chosen sample.

A variety of networks beyond the household has been used in studies of persons with various traits. Several are briefly summarized below:

Investigator	Network(s)	Rare Trait
Nathan (1976); Nathan et al.(1977) Levy (1977)	Offspring; siblings	marriages births
Rothbart et al. (1982)	siblings; aunts/uncles	Vietnam veterans
Sirken et al. (1980); Czaja et al. (1982); Czaja et al. (1986)	siblings; offspring	cancer patients
Bergsten and Pierson (1982)	immediate neighbor	gardeners using sewage sludge
Czaja (1988)	close friend; relative	crime victims

Sudman (1986)	parents;
	offspring;
	siblings;
	neighbors;
	close friends
		missing children

These applications of network sampling have either raised or addressed several fundamental questions related to the feasibility of this design strategy. How much do those who act as informants know about their networks? How much more effectively do network samples discover persons with rare traits than conventional samples? If the object is to interview all with the trait, how complete is the tracking information given by the informant? And finally, how successful will one be in contacting and interviewing members of the network with the trait? This paper addresses these questions within the context of a survey whose use of network sampling was intended to interview by telephone a sample of women early in pregnancy.

2. METHODS

In 1988, a grant was awarded to the UNC School of Public Health by the Association of Schools of Public Health and the Centers for Disease Control (CDC) to identify pregnant women to be used in a subsequent study on morbidity and health care utilization during pregnancy. Since pregnancy is a rare event in the general population, network sampling was tested as one of several strategies for identifying candidates for the subsequent study. The four networks tested for their feasibility in locating pregnant women and for the quality of information obtained through them were the household, sisters, building and religion. Since in network sampling the informants are asked about themselves as well as about members of their networks, each informant was considered a member of each network. The household network included all initial informants as well as women aged 15 to 49 living within their household. The sister and religion networks also included all initial informants as well as women aged 15 to 49 who were sisters of the initial informants or who belonged to the same church or synagogue. The building network was the same as the household network if the initial respondent lived in a single unit dwelling. Otherwise, the building network included the initial respondent as well as women aged 15 to 49 living in the same multiple-household building. Only women aged 18 to 49 were considered eligible for follow-up contacts.

The network survey to identify pregnant women for the health care utilization study was conducted by the spring 1988 BIOS 164 class under the direction of Dr. William Kalsbeek. The target population consisted of women aged 18 to 49 living in a five county area of central

North Carolina, and pregnant women aged 18 to 49 who were identified by one of the above as members of their household, sister, building or religious network. Telephone prefix information and a modified version of the Waksberg (Waksberg, 1978) approach for random digit dialing were used to obtain a sample of 200 primary sampling units (PSUs). The goal was to select five residential numbers (with an eligible woman living at each selected household) from each of the 200 PSUs. Some of the clusters (the PSUs) fell short of this goal resulting in a final sample of 869 initial respondents (key informants).

A cluster screening form and four questionnaires were developed by the class to accomplish the field testing. The cluster screening form was used to determine which 200 clusters would ultimately by included in the survey. For a cluster to be accepted, the first phone number chosen at random from within the cluster had to be a residential phone number with a woman aged 18 to 49 living in the household. Once the 200 clusters were accepted, the actual survey could begin. Three questionnaires were used to obtain pregnancy and demographic information. The Household Screening Questionnaire (HSQ) was the instrument used when key informants within the 200 clusters were initially contacted. This instrument solicited information concerning the size of the woman's networks, her knowledge about pregnancies within these networks, her interest in knowing about pregnancies, demographic information and tracking information to be used in contacting identified pregnant women. The Informant Pregnancy Questionnaire (IPQ) was administered to pregnant key informants. Further background information was obtained along with information concerning their current pregnancy and their feelings about telling others in the specified networks about their pregnancy. The Network Pregnancy Questionnaire (NPQ) was administered to pregnant women identified by the key informants as network members. It was similar to the IPQ in information solicited and question structure. Both pregnancy questionnaires attempted to obtain impersonal information first and gradually lead into the more sensitive issues concerning pregnancy history.

Analysis in much of this study required weights to account for the relative likelihood of appearance in the sample. A two-step process was followed for each of three sets of weights. A raw weight was first produced from a measure of the selection probability and then this weight was ratio-adjusted to the distribution of the population using the 19 telephone exchanges of the survey population as adjustment cells.

3. FINDINGS

For network sampling to be successful in locating women early in pregnancy, network members must be willing to tell others in the network about their pregnancy and the key informants must be aware of this information. Table 1 compares the proportion of key informants in each network who responded they were somewhat likely or very likely to find out about network members who were early in pregnancy. Since the informant's knowledge about their own pregnancy was not of interest, women were asked these questions only if there was more than one (including herself) member within the particular network of interest. The table also includes interest level as a subdomain. It seems reasonable to assume that women interested in learning about pregnancies are more likely to find out about them than uninterested women.

Women belonging to the household and sisters networks responded they were most likely to know about early pregnancies with proportions of .94 and .92 respectively. Women in the building network perceived themselves as being least likely of the four networks to know about early pregnancies. As expected, the table also indicates that for all but the building network, a greater proportion of women were likely to know about pregnancies if they were interested in learning about them than if they were uninterested.

As previously mentioned, for this design to be useful, network members must be willing to tell others in their network about their pregnancy. Pregnant key informants and women brought into the study as network members were asked about their willingness to tell others in the four networks about their pregnancy. Table 2 compares the proportion of women in each network who said they were willing to tell everyone in that network about their pregnancy. The marriage subdomain is included as one factor which may influence a woman's willingness to tell others. The household and sisters networks had the greatest proportion of women willing to tell others about their pregnancies with proportions of .81 and .83 respectively. Women were least likely to tell others in their religion network. Not surprisingly, married women were more likely to tell others in each of the networks about their pregnancy than were unmarried women.

Finally, since it was of interest to find women early in pregnancy, the women were also asked at what point they told others in each of the networks about their condition. Table 3 compares the average number of weeks when the women told others in each network. Women told others in their household sooner than any other network with an average time of 4.4 weeks. Women in the sisters and building networks also told relatively early with averages of 5.5 and 4.9 weeks respectively. Women in the religion network told later than the others with an average time of 8.2 weeks.

One key objective of the study was to evaluate the effectiveness of network sampling in locating members of a rare population (pregnant women) who would not otherwise be located through the use of a conventional survey design. One measure of this effectiveness was to compare the number of pregnant women identified through the household (conventional design) with the number identified through the use of networks. The average number of pregnant women identified per informant within each network and for all networks combined (combined network) can be found in Table 4. Since identification did not necessarily result in a completed interview, the average number of final interviews obtained per informant is also included in the table for the individual and combined networks.

By far the greatest number of identifications were obtained through the religion network with an average of 1.39 pregnant women identified per informant. The sisters network was next with an average of .10. The building and household networks were not far behind with averages of .08 and .05 respectively. Although at first glance the religion network appears far superior to the others in its effectiveness to identify pregnant women, this success is tempered by the high attrition that occurs between the initial identification and final interview. An average of only .09 final interviews per informant was obtained for the religion network. This was not much greater than the .05 obtained by each of the household and building networks which had the least number of final interviews per informant. It should by noted that the combined sample resulted in slightly more than twice the number of final interviews per informant than the household network considered alone. It also had considerably more attrition between the initial identification and final interview.

As was made clear in the last section, identification of a pregnant woman did not necessarily translate into a completed interview. For an interview to be obtained, contact had to be established between the interviewer and the identified population member. For this contact to occur, the interviewer required enough information to call the identified member.

All key informants were asked the name, phone and city of residence of any network members they identified as being pregnant. In Table 5, the number of identified pregnancies and completed interviews are stratified by the completeness of the tracking information. It appears that one likely reason for the high attrition seen in the religion network is the lack of useful tracking information obtained for a large proportion of identified pregnancies.

Table 5 also includes the ratios of completed interviews to total identified pregnancies stratified by completeness of tracking information. These data indicate that once a pregnancy is identified, the greatest success in obtaining an interview occurs when complete name phone and residence information is provided. Success is also high when complete phone and residence information is given. The least success occurs when no phone information is provided.

It must be noted that the total number of pregnancies appearing for the religion network are less than those actually reported. The questionnaire provided space for information on only nine subjects. Consequently, an informant may have said she knew of more than nine pregnant women but tracking information was only solicited for nine. No informant gave more than a residence for nine subjects. Consequently, it is assumed that useful tracking information was unavailable for the 361 women not appearing in the table.

Since the quantity of tracking information appeared to be a major factor in obtaining an interview once a pregnancy was identified, it was of interest to determine which network produced the most complete information. The weighted proportion of identified women for which complete name, phone and residence information was given can be found in Table 6. The proportion for which at least a phone number was given can also be found in this table. With proportions of .55 and .52, the most complete information was provided for women identified through the household and sisters networks respectively. Considerably less complete information was provided for women identified through the religion network. When phone information was considered separately, the household and sisters networks were again superior in the amount of information provided with proportions of 1.00 and .81. The building network did not lag far behind with 73% of the identified women having phone information provided for them. Again, the quantity of phone information provided by the religion network was considerably less than for the other networks. It should be noted that the proportions appearing in this table for the religion network are inflated due to the previously mentioned missing tracking information.

Following the acquisition of useful tracking information, the final step was to make a successful contact and obtain an interview. A contact was considered successful if someone at the contacted number verified it as being the number through which the identified pregnant woman could be reached. To gauge our success in contacting identified women and in obtaining interviews, the proportion of identified pregnancies which resulted in a successful contact or interview was calculated for each network. These values, which can be found in Table 7, indicate that the most effective network in terms of contacts made and interviews obtained was the household network. Approximately 98% of pregnancies identified through the household resulted in both a successful contact and interview. A possible explanation for this success is that 43 of the 46 pregnant respondents identified through the household were key informants who initially agreed to complete the Household Screening Questionnaire (HSQ). Administration of the pregnancy questionnaire did not require an extra call to locate the respondent or to request completion of the questionnaire since it was administered immediately following completion of the HSQ. This call was required for respondents located through the other networks. Identification through the sister or building networks resulted in successful contacts 72 or 78% of the time and in interviews 72 or 76% of the time, respectively. The religion network and the combined sample lagged far behind with 7 to 9% success rates for contacts and interviews. It should be noted that three of the successful religion contacts were considered out of scope since the women gave birth before contact was made. They were included in the successful contact proportion since our primary interest at this point was to determine the ability to contact a woman once she was identified through a particular network.

It was finally of interest to determine the proportion of contacts within each network which resulted in a successful interview. These proportions are presented in Table 8. Once contact was made, all networks resulted in a high proportion of successful interviews with the building and household networks having 100% success rates. The out of scope contacts were excluded from these calculations since no effort was made to obtain an interview once these women were contacted.

4. DISCUSSION

Several criteria for evaluating the overall effectiveness of network sampling and the relative efficiency of individual networks were outlined in the previous section. In addition, the mean squared error [MSE=Variance + $(\text{Bias})^2$] must be considered for each network. The variance component of the MSE is influenced by the number of discovered pregnancies that result in a successful interview as well as variation in sampling probabilities (resulting in variable sampling weights). As the number of successful interviews increase, the MSE decreases. Conversely, as sampling weights increase in variability the MSE increases. The size of the bias component is directly influenced by the percentage of identified pregnant women who become respondents in the study. High attrition between the time of identification and final interview leads to increased bias.

No network could be considered unequivocally superior when all criteria were considered. Although the religion network resulted in the greatest number of identified pregnancies per informant, it also had by far the greatest attrition among identified pregnancies. The household network had the lowest attrition but resulted in the least number of identified pregnancies and the lowest number of final interviews per informant. The sisters network resulted in two times the number of identified pregnancies than the household, but had an approximate 20% attrition rate whereas no attrition occurred within the household network. The building network had slightly greater attrition and slightly fewer identified pregnant women per informant than the sister network. Based on these results, the household and sister networks can be considered most effective in locating pregnant women and obtaining interviews. The religion and building networks can be considered least effective.

If the choice of study design is between the conventional household approach and network sampling, the conventional approach should be selected if the survey conditions are comparable to those in this study. Network sampling is considerably more complicated than conventional survey sampling in terms of questionnaire design, weight calculations and methods used to contact population members. With additional resources it seems likely we could have been more successful in extracting useful tracking information from key informants in the religion network. It then seems feasible that our study results could have favored network sampling over the conventional approach. It should still be noted however that neither approach appears too successful in locating women early (first trimester) in pregnancy. A large sample size may be required for any approach due to the rarity of the event.

If a multiple or combined network approach is considered, the yield of pregnancies per informant increases substantially over the single-network options, but the MSE must again be considered when deciding its relative merits. Overall attrition would be greater than the single-network options since it would approximate the weighted average of attritions obtained from the single-network options. Also, while the increased yield of pregnancies will result in a larger respondent sample size and resultant lower variance than any of the single-network options, the increase in the variance component due to variable sample weights will be much larger because of the considerable variation in the selection probabilities of identified pregnancies among the four networks. Since the net effect on the MSE cannot be determined from this study, the relative merit of the multiple-network option cannot be fully assessed.

In summary, network sampling did not achieve the desired aim of the Centers for Disease Control to locate a large sample of women early in pregnancy. However, there does appear to be potential for success if resources are increased to obtain better quality tracking information.

REFERENCES

Bergsten, J.W., and Pierson, S.A. (1982), "Telephone Screening for Rare Characteristics Using Multiplicity Counting Rules," Proceedings of the Section on Survey Research Methods, American Statistical Association, 145-150.

Czaja, R., and Blair, J. (1988), "Using Network Sampling for Rare Populations: An Application to Local Victimization Surveys," Survey Research Laboratory Working Paper, University of Illinois, Chicago.

Czaja, R.F., Snowden, C.B., and Casady, R.J. (1986), "Reporting Bias and Sampling Errors in a Survey of a Rare Population Using Multiplicity Counting Rules," Journal of the American Statistical Association, 81, 411-419.

Czaja, R., Warnecke, R.B., Eastman, E., Royston, P., Sirken, M., and Tuteur, D. (1982), "Locating Patients with Rare Diseases Using Network Sampling: Frequency and Quality of Reporting," Health Survey Research Methods, 311-324.

Kalton, G., and Anderson, D.W. (1986), "Sampling Rare Populations," Journal of the Royal Statistician Society, Serial A, 149, 65-82.

Levy, P.S. (1977), "Optimum Allocation in Stratified Random Network Sampling for Estimating the Prevalence of Attributes in Rare Populations," Journal of the American Statistical Association, 72, 758-763.

Nathan, G. (1976), "An Empirical Study of Response and Sampling Errors for Multiplicity Estimates with Different Counting Rules," Journal of the American Statistical Association, 71, 808-815.

Nathan, G., Schmelz, U.O., and Kevin, J. (1977), "Marriages and Births in Israel," Vital and Health Statistics, Series 2, No. 70.

Rothbart, G.S., Fine, M., and Sudman, S. (1982), "On Finding and Interviewing Needles in the Haystack: The Use of Multiplicity Sampling," Public Opinion Quarterly, 46, 408-421.

Sirken, M.G. (1970), "Household Surveys with Multiplicity," Journal of the American Statistical Association, 65, 257-266.

-----(1972a), "Variance Components of Multiplicity Estimators," Biometrics, 28, 869-873.

-----(1972b), "Stratified Sample Surveys with Multiplicity," Journal of the American Statistical Association, 67, 224-227.

Sirken, M., Royston, P., Warnecke, R., Eastman, E., Czaja, R., and Monsees, D. (1980), "Pilot of the National Cost of Cancer Care Survey," Proceedings of the Section on Survey Research Methods, American Statistical Association, 579-584.

Sudman, S. (1985), "Experiments in the Measurement of the Size of Social Networks," Social Networks, 7, 127-151.

-----(1986), "The Use of Network Samples in Estimating Incidence of Missing Children," Proceedings of the Section on Survey Research Methods, American Statistical Association, 159-163.

Sudman, S., and Kalton, G. (1986), "New Developments in the Sampling of Special Populations," Annual Review of Sociology, 12, 401-429.

Table 1 Weighted proportion of key informants who responded they were likely to know about network members who were early in pregnancy

	Household	Sisters	Building	Religion
	P (N)	P (N)	P (N)	P (N)
TOTAL SAMPLE	.94 (178)	.92 (568)	.49 (317)	.61 (432)
INTEREST LEVEL				
Interested	.95 (146)	.93 (472)	.47 (254)	.65 (352)
Uninterested	.88 (30)	.87 (94)	.56 (60)	.40 (78)

Note: Sample sizes given in parenthesis

Table 2 Weighted proportion of women who responded they were willing to tell all in network about their pregnancy

	Household	Sisters	Building	Religion
	P (N)	P (N)	P (N)	P (N)
TOTAL SAMPLE	.81 (74)	.83 (87)	.64 (81)	.30 (66)
MARITAL STATUS				
Married	.86 (67)	.84 (77)	.72 (72)	.33 (61)
Unmarried	.37 (7)	.73 (10)	.14 (9)	.01 (5)

Note: Sample sizes given in parenthesis

Table 3 Weighted average time when women told others in network about their pregnancy

	Household	Sisters	Building	Religion
	(73)	(81)	(69)	(63)
WEEKS	4.43	5.55	4.89	8.21

Note: Sample sizes given in parenthesis

Table 4 Weighted average number of identifications and interviews per informant

	HH	Sisters	Building	Religion	Combined Network
AVERAGE # OF PREGNANCIES IDENTIFIED PER INFORMANT	.05	.10	.08	1.39	1.46
AVERAGE # OF FINAL INTERVIEWS PER INFORMANT	.05	.08	.05	.09	.12

Note: Sample Size=869

Table 5 Profile of completeness of tracking information

PATTERN			HH	Sisters	Building	Religion	Combined Network
			# Completed Interviews				
Name	Phone	Residence					
1	1	1	25	39	26	58	73
1	1	0	0	2	0	2	4
0	1	1	20	21	21	21	23
0	1	0	0	0	1	1	1
1	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	1	0	0	1	1	1
0	0	0	0	0	0	0	0
			Total # Identified Pregnancies				
1	1	1	25	43	26	74	93
1	1	0	0	2	0	8	10
0	1	1	21	23	22	25	28
0	1	0	0	0	0	1	1
1	0	1	0	1	1	21	23
1	0	0	0	0	0	1	1
0	0	1	0	1	16	35	52
0	0	0	0	12	1	697	709
			# Completed Interviews/Total # Identified Pregnancies				
1	1	1	1	.91	1	.78	.78
1	1	0	0	1	0	.25	.40
0	1	1	.95	.91	.95	.84	.82
0	1	0	0	0	0	1	1
1	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	1	0	0	0	.03	.02
0	0	0	0	0	0	0	0

Key: 1-Complete information given
0-Complete information not given

Table 6 Weighted proportion of identified pregnant women for which complete tracking information was given and proportion for which at least phone was given

	HH	Sisters	Building	Religion	Multiple Network
	(46)	(82)	(65)	(862)	(917)
COMPLETE INFORMATION GIVEN	.55	.52	.41	.08	.10
AT LEAST PHONE GIVEN	1.00	.81	.73	.12	.14

Note: Sample sizes given in parenthesis

Table 7 Proportion of successful contacts and interviews obtained following pregnancy identification

	HH	Sisters	Building	Religion	Multiple Network
	(46)	(82)	(65)	(1223)	(1278)
# SUCCESSFUL CONTACTS	45	64	47	88	109
# CONTACTS/ TOTAL IDENTIFIED	.98	.78	.72	.07	.09
# SUCCESSFUL INTERVIEWS	45	62	47	83	102
# INTERVIEWED/ TOTAL IDENTIFIED	.98	.76	.72	.07	.08

Note 1: Total # identified pregnancies given in parenthesis
Note 2: Out of scope contacts included

Table 8 Proportion of successful contacts resulting in an interview

	HH	Sisters	Building	Religion	Multiple Network
	(45)	(64)	(47)	(85)	(106)
# SUCCESSFUL INTERVIEWS/ # SUCCESSFUL CONTACTS	1.00	.97	1.00	.98	.96

Note: # successful contacts given in parenthesis-- out of scope contacts excluded