# A SHORT HISTORY OF NETWORK SAMPLING

Monroe G. Sirken, National Center for Health Statistics
6525 Belcrest Road, Room 1000, Hyattsville, MD. 20782

**Key Words: multiplicity surveys, sampling history, survey design**

## Introduction

Network sampling and classical survey sampling differ with respect to the counting rule paradigm for linking population elements to the selection units at which they are countable in the survey [20]. Classical survey sampling uses **unitary counting rules**, such as de jure and de facto residence rules in household surveys, that seek to uniquely link each person to one and only household. Network sampling, on the other hand, seeks to capitalize on duplicate counting of population elements by using **multiplicity counting rules**, such as friendship and kinship rules in household surveys, that link the same person to multiple households of their friends or relatives.

Over the past thirty years, network sampling has improved survey design efficiencies particularly when classical sampling is infeasible or inefficient. During the 1960's, network sampling was applied in establishment surveys when unitary counting rules are difficult to define and execute because the same population elements appear inextricably linked to multiple establishments. During the 1970's, network sampling was fostered in household surveys of rare populations in which de jure residence rules are easy enough to define and execute, but the sampling error effects are often intolerably large. During the 1980's, network sampling was extended to rare population surveys in which de jure residence rules incur large measurement errors as well as sampling errors. I'll briefly discuss and illustrate each these network sampling applications.

During the 1990's, network sampling theory is being applied in population based establishment surveys [13, 23]. In these surveys, establishments that have transactions with persons enumerated in household sample surveys serve as sampling frames for establishment surveys. Population based establishment surveys provide a mechanism for integrating the sample designs of household and establishment surveys to produce statistics about transactions that people have with establishments . They are especially applicable when free-standing establishment frames do not exist or if available do not have good measures of establishment size.

## Establishment surveys in which unitary counting rules are difficult to apply

In the early 1960's, resolution of an estimation problem in a medical provider survey to estimate the prevalence of cystic fibrosis [7] ultimately led to the development of network sampling. Classical sampling estimation was not applicable in this survey because multiple medical providers often treated and hence reported the same cystic fibrosis patients. Birnbaum and Sirken [1] resolved the problem by proposing three unbiased estimators for medical provider surveys of rare disease prevalence in which multiple providers are eligible to report the same patients.

The three estimators proposed by Birnbaum and Sirken utilize information about the multiplicities of medical providers eligible to report the same patients in the survey. This information at typically collected when the providers report their patients in the survey. The estimators differ from each other with respect to the way the multiplicity information is used. The **multiplicity estimator,** for example, counts every patient, as many times as he or she is reported by a different medical provider in the sample survey, and weights each report by the inverse of the patient's multiplicity. On the other hand, the **Horvitz/Thompson network estimator** calculates the selection probabilities of every patient reported in the survey, and that requires counting the number of times the same patients are reported by different medical providers in the survey, and knowing the multiplicities of every reported patient.

Since the 1960's network sampling has been applied in many medical provider surveys and other kinds of establishment surveys. In these applications, the design strategy is to select the counting rules that minimize total survey errors [17]. Sometimes, field experiments are undertaken to investigate alternative counting rule options.

For example, Hendricks, Searles and Horvitz [6] compare efficiencies of alternative counting rules or associating farms and crop acreage with areal segments in agriculture sample surveys. Traditionally, agriculture surveys used the "headquarters" counting rule. This rule links each farm to one and only one area segment, namely the segment containing the farm's headquarters. Difficulties in defining farm headquarters, and in implementing this rule, led to testing the "weighted segment" rule. This rule links each farm to every area segment intersecting the farm's boundaries. Farms intersected by sample area segments were weighted by the fractions of the farms' land within the area segments. This **weighted multiplicity estimator** is an unbiased estimation procedure, when as in this example the sums of the fractional weights assigned each farms in intersected area segments equal unity.

In Hendricks' experiment, sampling errors associated with the headquarters were reduced by about 25 to 50 percent by area segment rule. Furthermore, interviewers misinterpreted and misapplied the weighted segment rule far less frequently than the headquarters rule.

### Household surveys of rare populations and events

Network sampling emerged as a distinct type of sample design during the 1970's when it was deliberately fostered as a design strategy in household surveys of rare populations that use **composite counting rules.** These rules have the property of linking rare persons to their own residences and to other residences of persons, with whom they have well defined relationships, such as relatives, friends, or neighbors.

Though composite counting rules enumerate more persons than de jure residence rules, sampling variances associated with composite counting rules are not necessarily smaller [14] because other factors are relevant including the extend of clustering, and variability in the multiplicities [15]. However, when composite counting rules link no more than one rare person to any household, they reduce variances associated with the de jure rule by a factor equal to the harmonic mean of the multiplicities of the rare population. If, for example, the composite rule links no more than one rare person to a household, and assigns all rare persons the same multiplicity, say s, classical sampling variance associated with the de jure rule is reduced by a factor of s.

On the other hand, reporting biases are often larger for composite counting rules than de jure residence rules because the latter involve collecting supplementary survey information. In compliance with both rules, households report their residents that have the rare attribute. In addition, in compliance with the composite rule, households report the multiplicities of their own household members having the rare attribute, and they serve as proxy respondents for non household persons to which they are linked by the composite rule, and report which of them have the rare attribute, and their multiplicities.

Because of the difference in the relative magnitudes of sampling errors and reporting bias associated with the composite and de jure counting rules, relative efficiencies of composite rules typically decrease with increasing sample size.

### Surveys of rare events

Nathan [10] and Nathan, Schmelz, and Kenvin [11] compared efficiencies of the de jure residence rule and a composite counting rule in a natality household sample survey to estimate the number of births. The de jure residence rule links births to residences of their mothers, and the composite counting rule links births to residences of infants' mothers, and maternal grandmothers and aunts. The experiment was embedded in the Israel Labor Force Survey during the first quarter of 1974, and respondents retrospectively reported births that occurred during 1973. Reporting bias was evaluated by a quality check survey that interviewed residences of maternal grandmothers and aunts when mothers in sample households reported births, and interviewed residences of mothers when maternal grandmothers or aunts in sample households reported births of grandchildren or nieces or nephews.

The variance of the survey estimate of the number of births associated with the de jure residence rule is reduced by almost fifty percent by the maternal grandmother/aunt composite counting rule. The net reporting bias, however, is slightly smaller for the de jure rule, +0.08 percent, than for the composite rule, +0.59 percent. (Absolute values of the under reporting and over reporting biases are substantially larger, 9.14 percent and 9.73 percent, respectively for the composite rule, than the 2.04 and 2.12 percent, respectively for the de jure rule). The composite counting rule is more efficient then the de jure rule to sample sizes up to about 4,400 households for total births, and up to a considerably larger number of households for estimates of births by demographic and geographic subdomains.

Compared to the official demographic estimate, the de jure rule and the composite rule undercounted the number of Israel births in 1973 by almost 7 percent. Dual system estimators are often used to adjust for under enumeration in surveys and censuses. The **dual system network estimator** [2, 18] is the network sampling version of the classical sampling dual system estimator [8]. The former uses a **disjoint counting rule** that links events to households by two independent counting rules (e.g. a de jure residence rule and a kinship rule), such that the same events are not linked to the same households by both counting rules. The number of events that would be reported by both counting rules is estimated by conducting quality check surveys that interview households eligible to report events by the counting rule alternative to the one by which those same events were originally reported in the survey.

To my knowledge, dual system network estimation procedures have not been field tested even though they appear to have considerable potential for improving design efficiency of quality check surveys such as the post enumeration survey (PES) to evaluate completeness of enumeration in decennial censuses. Small pilot tests of PES designs based on network sampling were undertaken in preparing for the 1980 U.S. Census of Population and Housing [9, 22]. However, these PES pretests of network sampling were undertaken prior to the development of dual system network estimation.

**Surveys of rare populations**

Czaja, Snowden, and Casady [4] compare efficiencies of cancer prevalence estimates associated with the de jure residence rule linking cancer patients to their own residences and a composite counting rule linking cancer patients to their own residences and those of their children. The experiment involves 530 households of which 325 were households of cancer patients selected from cancer registries of two Illinois hospitals, and 205 were Illinois households of children of the cancer patients. Estimates of under reporting biases are obtained by matching the cancer patients reported in the survey with the two cancer registers [3, 21].

Sampling variance of the cancer prevalence estimates based on the de jure residence rule is reduced by about 50 percent by the children composite counting rule. Under reporting biases are 11.0 percent and 14.3 percent respectively for de jure residence rule and the children composite counting rule for all patient domains. For combined cancer sites, the children composite rule is more efficient than the de jure rule for sample sizes up to about 4200 households, and up to substantially larger sample sizes for specific cancer sites.

Under reporting bias varies by sex, age and race of patient. For example, about 3 percent and 7 percent respectively of white female cancer patients are not reported at residences of patients and children. The comparable under reporting biases for white male cancer patient are 12 percent and 16 percent, respectively. For white female cancer prevalence, the children composite rule is more efficient than the de jure rule for sample sizes up to about 80,000 households.

**Household surveys of rare and elusive and/or sensitive populations**

If the rare population is also elusive or if the rare attribute is a sensitive one, sample survey estimates associated with the de jure residence rule are vulnerable to large reporting biases as well as large sampling errors. Under either of these circumstances, network sampling offers options that may be more efficient than classical sampling. For example, the likelihood of enumerating elusive populations, such as migrants, nomads and the homeless, may be better if they are enumerated at the fixed residences of knowledgeable close associates, such as relatives and friends, than if enumerated at elusive persons' own residences. Similarly, the likelihood of enumerating populations with sensitive attributes may be better if they are enumerated at residences of friends and relatives since that venue provides greater response anonymity, than the residences of persons with the sensitive attributes.

**Surveys of elusive populations**

Decedents represent an elusive population in retrospective mortality household sample surveys using the de jure residence rule. Institutional deaths are missed because they aren't linked to any households by the de jure residence rule. Noninstitutional deaths are often missed because the decedents' former households dissolve before the surveys are conducted.

Sirken and Royston [24] compare efficiencies of mortality surveys using a de jure residence rule linking decedents to their noninstitutional places of residence at death, and a composite counting rule linking decedents to their former residences and to residences of surviving spouses, siblings and children. In the survey experiment, which was conducted during 1975, respondents retrospectively reported deaths that occurred during 1974. Interviews were conducted at former noninstitutional residences of decedents, and at the current residences of the decedents' surviving close relatives for a sample of several hundred registered deaths that occurred in North Carolina during 1974. Under reporting bias was assessed by matching the deaths reported in the survey experiment with files of registered deaths in North Carolina. Although decedents at all ages are included in the experiment, the findings reported here refer to decedents, 65-84 years of age.

The kinship composite counting rule is uniformly more efficient than the de jure residence rule for estimating the number of noninstitutional deaths and even more efficient for estimating the combined number of institutional and noninstitutional deaths. Sampling variance of the de jure residence rule was reduced by about 75 percent by the kinship composite rule. The fraction of missed institutional and noninstitutional deaths was reduced by almost one half, from 29 percent, to 15 percent by the composite counting rule. The de jure rule missed all institutional deaths, which represented about 22 percent of all deaths, and the composite counting rule missed about a third of the institutional deaths. Both counting rules failed to enumerate about 7 percent of the noninstitutional deaths [19].

### Surveys of sensitive populations

Rittenhouse and Sirken [12] compare efficiencies of the de jure residence rule linking heroin users to their own residences, and a composite counting rule linking heroin users to their own residences and residences of their close friends. The experiment was embedded in a half sample (2250 household) of the 1977 National Survey on Drug Abuse.

Estimates of lifetime heroin use prevalence are substantially higher for the friends counting rule (5.8 percent) than for the de jure residence rule (1.3 percent). However, sampling variances are almost twice as large for the composite rule than the de jure rule. This somewhat surprising finding is due to extensive clustering of heroin users within friendship networks and considerable variability in the heroin users' multiplicities. Nevertheless, the composite friends rule is far more efficient than the de jure rule assuming the validity of the composite rule's higher lifetime heroin use prevalence estimate, which was in close agreement with expert opinion on lifetime heroin use prevalence during 1977.

In this experiment, about ten percent of the respondents reported close friends that were heroin users, but almost a third of them failed to report the multiplicities of their heroin user friends [5]. Therefore, the lifetime heroin prevalence was also estimated by the **hybrid network estimator** [23]. This network estimator utilizes all reports of heroin use reportable in compliance with the friends composite rule, and utilizes multiplicities when heroin use is self reported, but does not utilize multiplicities when heroin users are reported by friends. The hybrid network estimate of lifetime heroin prevalence is 2.8 percent, or about midway between the multiplicity estimate and the estimate based on the de jure rule. Sampling errors are roughly 1.5 to 3 times larger for the hybrid estimate than for the multiplicity estimate.

My curiosity about the potential utility of the friends counting rule to estimate heroin use was initially aroused by survey estimates of illicit substance use that are based on respondents' reports of the percentages of their friends using illicit substances. Illicit substance use estimates are substantially higher for reports of percentage friends heroin use than self reports of heroin use. In the 1974 Michigan Survey of Drug Abuse, for example, prevalence estimates of about a half dozen illicit drugs are 50 to 200 percent higher based on reports of percentage friends use than for self reports of substance use [16]. On the other hand, the prevalence estimates of several prescribed drugs, including amphetamines, narcotics and tranquilizers, are mostly higher based on self reports than reports of percentage friends use. The contrasting effects of the counting rules on the prevalence of prescribed drugs and illicit drugs, suggest that anonymity of response provided by the friends rule enhanced the likelihood of truthful response about illicit drug use.

From a statistical viewpoint, the Michigan Survey findings are quite puzzling: averaging percentages of friends reported as drug users, the estimator in the Michigan Survey, is a biased estimator. It would be an unbiased estimator if and only if friends of illicit drug users form closed networks in which friendship ties between drug users and friends are reciprocal and neither drug users nor their friends have other friendship ties. There are multiplicity rules that form closed networks, but evidence is lacking that friends of drug users is one of them. For example, the sibling counting rule and the maternal or paternal first cousin counting rules are transparent examples of **closed counting rules** that form closed networks.

### The future of network sampling

The future of network sampling is interdisciplinary survey methods research. Network sampling research intersects the cognitive, behavioral, and statistical sciences. For example, fundamental knowledge about information networks linking relatives and friends is critical in designing surveys based on network sampling, and knowledge gained about the robustness of these information networks from survey applications of network sampling is potentially valuable in sociological research. Also, fundamental knowledge about cognitive aspects of information processing is essential in designing complex questionnaires for household surveys using network sampling, and knowledge gained by observing respondents respond to network survey questionnaires is potentially valuable in cognitive science in stimulating new areas of cognitive research (Sirken and Schechter, in press).

**References**

[1] Birnbaum, Z.W. & Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates, *Vital and Health Statistics*, PHS Publication No. 1, Series 2, No. 11. US Government Printing Office, Washington.

[2] Casady, R.J., Nathan, G. & Sirken, M.G. (1985). Alternative dual system network estimators, *International Statistical Review* **53**, 183-197.

[3] Czaja, R., Warnecke, R.B., Eastman, E., Royston, P., Sirken, M. & Tuteur, D. (1984). Locating patients with rare diseases using network sampling: frequency and quality of reporting, in *Proceedings of the Fourth Conference on Health Survey Research Methods.* Public Health Publication No. 84-3346, National Center for Health Service Research, US Department of Health and Human Services, pp. 311-324.

[4] Czaja, R.F., Snowden, C.B. & 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.

[5] Fishburn, P.M. (1979). Heroin estimator development. Unpublished report submitted to the National Institute On Drug Abuse.

[6] Hendricks, W.A., Searles, D.T. & Horvitz, D.G. (1965). A comparison of three rules for associating farms and farmland with sample area segments in agriculture surveys, in *Estimation of Areas in Agricultural Statistics, Food and Agriculture*. Organization of the United Nations, Rome, pp. 191-198.

[7] Kramm, E.R., Crane, M.M., Sirken, M.G. & Brown, M.L. (1962). A cystic fibrosis pilot study in three New England states. *American Journal of Public Health* **52**, 2041-2057.

[8] Marks, E. (1978). The role of dual system estimation in Census evaluation, in *Development in Dual System Estimation of Population Size and Growth*, The University of Alberta Press. Edmonton, Alberta, Canada.

[9] Marks, E. & Ockay, C. (1978). A model for network (multiplicity): estimation of census under coverage, in *American Statistical Association 1978 Proceedings of the Section on Survey Research Method.* American Statistical Association, Alexandria.

[10] 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, pp 803-815.

[11] Nathan, G., Schmeltz, O. & Kenvin, J. (1977). Multiplicity Study of Marriages and Births in Israel, *Vital and Health Statistics*, Series 2, No. 78. DHEW Publication No. (PHS) 79-1352. US Government Printing Office, Washington.

[12] Rittenhouse, J.D. & Sirken, M.G. (1981). A note on networks, nominations, and multiplicity, as contributory to heroin estimation. *Administrative Report*, National Institute of Drug Abuse, Department Health and Human Services, Washington.

[13] Shimizu, I. & Sirken, M.G (1998). More on population based establishment surveys, in *American Statistical Association 1998 Proceedings of the Section in Society Research Methods*, in print.

[14] Sirken, M.G. (1970). Household surveys with multiplicity, *Journal of the American Statistical Association* **65**, 257-266.

[15] Sirken, M.G. (1972). Variance components of multiplicity estimators, *Biometrics* **28**, 869-873.

[16] Sirken, M. G. (1975). Evaluation and critique of household sample surveys of substance abuse, in *Alcohol and Other Drug Use and Abuse In the State of Michigan*, Michigan Department of Public Health, Lancing pp 1-35.

[17] Sirken, M.G. (1975). The counting rule strategy in sample surveys, in *American Statistical Association 1975 Proceedings of the Section on Social Statistics*, American Statistical Association, Alexandria, pp. 119-123.

[18] Sirken, M.G. (1979). A dual system network estimator, in *American Statistical Association 1979 Proceedings of the Section on Survey Research Methods.* American Statistical Association, Alexandria, pp. 340-342.

[19] Sirken, M.G. (1983). Handling mission data by network sampling, in *Incomplete Data in Sample Surveys, Part III*, Vol. 2. Academic Press, New York, pp. 81-90.

[20] Sirken, M. G. (1998). Network Sampling, in *Encyclopedia of Biostatistics*, Volume 4. John Wiley and Sons, pp 2977-2986.

[21] Sirken, M., Royston, P., Warnecke, R., Eastman, E., Czaja, R. & Monsees, D. (1980). Pilot of the national cost of cancer care survey, in *American Statistical Association 1980 Proceedings of the Section on Survey on Survey Research Methods*, pp 579-584.

[22] Sirken, M.G., Graubard, B.L., & LaValley (1978). Evaluation of Census population coverage by network surveys, in *American Statistical Association 1978 Proceedings of the Section on Survey Research Methods*, pp 239-243.

[23] Sirken, M.G. & Nathan, G. (1988). Hybrid network estimators, in *American Statistical Association 1988 Proceedings of the Section on Survey Research Methods.* American Statistical Association, Alexandria, pp. 459-461.

[24] Sirken, M.G. & Royston, P.N. (1976). Design effects in retrospective surveys, in *American Statistical Association 1976 Proceedings of the Section on Social Statistics*, pp 773-777.

[25] Sirken, M.G., Shimizu, I. & Judkins, D. (1995). Population based establishment surveys, in *American Statistical Association 1995 Proceedings of the Section on Survey Research Methods.* American Statistical Association, Alexandria, pp. 470-473.