SURVEY STRATEGIES FOR ESTIMATING RARE HEALTH ATTRIBUTES

MONROE G. SIRKEN
NATIONAL CENTER FOR HEALTH STATISTICS

1. Introduction

Estimation of the incidence and prevalence of rare health attributes in the population is one of the most difficult and persistent methodological problems in the national program for producing health and vital statistics. Speakers at this symposium have called attention to this methodological problem with respect to planning epidemiological studies of pollutant effects. They pointed out that many of the most serious health conditions such as congenital malformations, infant deaths, and numerous severe chronic diseases in which pollutants have been implicated or suspected, affect relatively small numbers of persons.

One objective of this paper is to describe the sample survey methods that have been used by the National Center for Health Statistics (NCHS) to produce national statistics for health conditions with low rates of prevalence and for vital events with low occurrence rates. Since different data systems have evolved in this country for producing vital statistics and for producing morbidity statistics, the methodological problems associated with estimating rare vital events are somewhat different than those associated with rare health conditions and the methods of dealing with these problems have been somewhat different also. Therefore, the matter will be discussed separately for the two data systems.

Another objective of this paper is to describe a new type of estimator that is currently being investigated by NCHS. The estimator is being tested in sample surveys of providers of health services to estimate rare health conditions and in household sample surveys to estimate rare vital events.

2. Rare vital events

National birth and death statistics are predominately by-products of the birth and death registration systems. Vital statistics are derived from the items of information reported on the records of registered births and deaths. Since national vital statistics are based on 100 per cent of the nearly two million deaths registered annually and on a 50 per cent sample of the nearly four million annually registered births, estimating the number of rare vital events in terms of the demographic and medical variables on the records does not present a problem.

However, the number of statistical items of information recorded on the vital records is restricted and the items tend to refer to variables that are closely tied to the date that the event occurred. Thus, the death record identifies the decedent's usual place of residence on the date of his death but it does not record his prior places of residence that would be required for ascertaining the decedent's longer term exposure to air pollutants. The number and kinds of items on the vital records are limited because the records are primarily legal rather than statistical records, and they are often used for personal identification purposes. Since 1900 the revisions in the standard certificates of birth and death have been minor [8] despite a dramatic increase in the demand to expand the scope of vital statistics. Consequently, innovative methods are needed to estimate the occurrence of vital events that are not defined in terms of the variables recorded on the vital records themselves.

Two kinds of sample survey techniques have been used to expand the scope of national vital statistics: (1) sample surveys linked to vital records and (2) dual sample surveys. The first is a method to estimate the frequency of rare vital events and the second is a method to estimate the occurrence rates of the rare events.

The design of sample surveys linked to vital records has been described in detail elsewhere [9], [19]. In these surveys, the files of the vital records serve as the sampling frame. They are excellent sampling frames because the files are virtually complete and each record in the files contains information which may be used in the sample selection and estimation process to improve the efficiency of the survey. Furthermore, the vital records provide names and addresses of persons and institutions that serve as the primary sources of information collected in the surveys. For example, the death record identifies the medical certifier of the causes of death, the hospital or institution in which the death occurred, and the "death record informant" who is typically a close relative of the deceased person. The birth record identifies the medical attendant at birth, the hospital in which the birth occurred, and the parents. Effective data collection procedures have been developed [16], [17] which depend primarily on mail surveys with provision as required for personal interview follow-up on subsamples of nonrespondents.

The program of conducting *ad hoc* surveys linked to birth and death records began about 15 years ago [18]. The national statistical program for conducting these surveys on a continuing basis was started about a decade ago [13]. Some examples of national statistics derived from these programs are listed below:

(a) A national survey linked to birth records produced statistics on medical irradiation exposure of the mother during pregnancy [10].

(b) A national survey linked to infant death records produced infant death statistics by fertility history of the mother, and by birth weight of the infant [11].

(c) A national survey linked to records of adult decedents produced mortality statistics by residence history and by smoking history of the decedents [3].

In dual sample surveys, separate surveys are conducted to estimate the

numerators and the denominators of vital rates. The numerator of the rate is the number of vital events that occurred during a specified calendar period. The denominator of the rate is the size of the population exposed to the risk of the event during the specified calendar period. The numerator is estimated by conducting a survey linked to vital records. The denominator is estimated by conducting a survey of the exposed to risk population which is usually a household sample survey although it may be a survey linked to vital records. A few examples of the kinds of vital rates that have been derived from dual sample surveys are listed below:

(a) Infant death rates by socioeconomic status. The numerators were based on a survey linked to infant death records and the denominators were based on a survey linked to birth records [11].

(b) Death rates by smoking history and residence history. The numerators were based on a survey linked to death records and the denominators were based on a national household sample survey [3].

3. Rare health conditions

Morbidity statistics are collected in the National Health Survey [12], which comprises a family of sample surveys in which each survey produces distinctive health and related population statistics. For example, the Health Interview Survey (HIS) produces statistics on the social dimensions of illness and the impact of morbidity on the population. On the other hand, the Hospital Discharge Survey (HDS) produces statistics on hospital utilization. Estimates of rare health conditions derived from these surveys are subject to very large sampling errors because the complex designs of the surveys involve relatively small samples. The HIS is based on interviews conducted in a national sample of about 35,000 households annually. The HDS is based on information abstracted from hospital records for a national sample of about 200,000 discharges annually or less than 0.5 per cent of the discharges from short stay hospitals.

The *ad hoc* survey of medical processes is one possible solution to the problem of designing sample surveys to estimate the number and characteristics of persons with specified rare health conditions. In this type of survey, listings of medical sources serve as the sampling frame. The cystic fibrosis survey was a prototype *ad hoc* survey of hospitals and physicians to estimate the incidence, prevalence, and case fatality rates of diagnosed cases of cystic fibrosis, a relatively rare genetic disease affecting roughly 1 in 2500 live births. A sample of medical sources stratified by size of hospital and specialty of physician was selected. In the mail survey, every sample source reported the patients with the disease that it had treated.

The survey presented an interesting estimation problem because cystic fibrosis patients are frequently treated by more than one medical source. Since medical sources were the sampling elements, patients had different probabilities of being reported in the survey, the probabilities being proportional to the

number of medical sources in the population who treated them for the disease. Work on this problem led to the development of estimators, which utilize ancillary information on the extent of multi-reporting of the patients by different medical sources.

In the cystic fibrosis survey, the extent of multiple reporting of patients was determined on the basis of the following types of auxiliary data that were collected for each patient reported by a sample source:

(a) The sample source who reported a patient identified other medical sources known to him who also treated his patient.

(b) Nonsample sources who were reported by sample sources as having treated their patients were added to the survey to verify that they had treated these patients and to determine if there were other medical sources who had also treated these patients.

Actually three different unbiased estimators were derived for the cystic fibrosis survey [7] and other estimators have been proposed [2], [5] for dealing with the problem. The cystic fibrosis estimators differed in the way that they utilized the information collected in the survey about multiple reporting of the same patient by different medical sources. The simplest of the three is the multiplicity estimator.

4. Comparison of multiplicity and conventional estimators

There are I_α, $\alpha = 1, \cdots, N$, individuals in the population who have a specified health condition. The problem is to estimate N. There are S_i, $i = 1, \cdots, L$, medical sources from which a sample S_j, $j = 1, \cdots, \ell$, sources is selected without replacement. The estimator of N is

$$\hat{N} = \frac{L}{\ell} \sum_{j=1}^{\ell} \lambda_j \tag{1}$$

where λ_j represents the information about individuals with the health condition reported in the survey by the jth sample source.

In the conventional survey a counting rule is adoped such that each patient is uniquely reported by a single source. For example, in the cystic fibrosis survey such a rule might state that "each patient is reported in the survey by the one source that has the major responsibility for treating the disease." Under conventional conditions

$$ {}_c\lambda_j = \sum_{\alpha=1}^{N} {}_c\delta_{\alpha,j}, \tag{2}$$

$$ {}_c\delta_{\alpha,j} = \begin{cases} 1 & \text{if } S_j,\ j = 1, \cdots, \ell, \text{ reports } I_\alpha \text{ in the conventional survey} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

According to the conventional rule

$$\sum_{i=1}^{L} {}_c\delta_{\alpha,i} = 1 \tag{4}$$

since one and only one source in the population reports I_α. Thus the conventional estimator of N is

$$\hat{N}_c = \frac{L}{\ell}\sum_{j=1}^{\ell} {}_c\lambda_j = \frac{L}{\ell}\sum_{j=1}^{\ell}\sum_{\alpha=1}^{N} {}_c\delta_{\alpha,j}. \tag{5}$$

In the multiplicity survey (that is, the survey using the multiplicity estimator) a multiplicity rule is adopted such that each patient is reported by at least one source. For example, in the cystic fibrosis survey, the rule stated that "each patient is reported by every medical source that ever treated him for the disease." Under these circumstances

$$ {}_m\lambda_j = \sum_{\alpha=1}^{N} \frac{{}_m\delta_{\alpha,j}}{s_\alpha} \tag{6}$$

where

$$ {}_m\delta_{\alpha,j} = \begin{cases} 1 & \text{if } S_j,\ j = 1, \cdots, \ell, \text{ reports } I_\alpha \text{ in the multiplicity survey} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and

$$\sum_{i=1}^{L} {}_m\delta_{\alpha,i} = s_\alpha = \text{number of sources in the population reporting } I_\alpha. \tag{8}$$

The multiplicity estimator of N is

$$\hat{N}_m = \frac{L}{\ell}\sum_{j=1}^{\ell} {}_m\lambda_j = \frac{L}{\ell}\sum_{j=1}^{\ell}\sum_{\alpha=1}^{N} \frac{{}_m\delta_{\alpha,j}}{s_\alpha}. \tag{9}$$

Some features of this estimator are particularly noteworthy:

(a) The s_α are needed *only* for I_α that are reported by sample sources.

(b) The survey procedure for determining the s_α does not necessarily require a survey of nonsample sources if this information can be reported by the sample sources.

(c) There is no need to match the patients reported by different sample sources in order to eliminate duplicate reports of the same patient.

(d) The conventional as well as the multiplicity estimates can be derived from the multiplicity survey if the multiplicity rule incorporates the conventional counting rule.

Both $\hat{N}_m$ and $\hat{N}_c$ are unbiased estimators of N provided that every patient is reported by one and only one source in the conventional survey and by at least one source in the multiplicity survey. The multiplicity estimator, however, involves the collection of more data from the sample sources in the survey. Clearly, the number of patients reported per source is greater in the multiplicity than in the conventional survey. Thus,

$$\frac{1}{L}\sum_{\alpha=1}^{N}\sum_{i=1}^{L} {}_m\delta_{\alpha,i} = \frac{1}{L}\sum_{\alpha=1}^{N} s_\alpha \geqq \frac{1}{L}\sum_{\alpha=1}^{N}\sum_{i=1}^{L} {}_c\delta_{\alpha,i} = \frac{N}{L}. \tag{10}$$

In addition, ancillary information is collected in the multiplicity survey in order to determine the s_α of each I_α reported by a sample medical source. This

ancillary information is not required in the conventional survey because according to the conventional estimator I_α, $\alpha = 1, \cdots, N$, is linked to a single source.

It has been shown [14] that $\hat{N}_m$ is not necessarily a more efficient estimator than $\hat{N}_c$. Which of the estimators has the smaller sampling variability depends on the particular counting rules adopted in the conventional and multiplicity surveys. The statistician, however, decides which particular counting rule, either conventional or multiplicity, is adopted in each survey. Obviously, the multiplicity estimator should be used selectively in those surveys where it is believed that a multiplicity rule produces a more efficient estimate than a conventional counting rule.

Some guidelines have been developed for selecting efficient multiplicity rules in surveys based on simple random sampling [14] and stratified sampling [15]. Some conditions for making the multiplicity rules more efficient than conventional counting rules are more likely to be met in surveys of rare attributes than in other surveys. For example, if the multiplicity rule satisfies the condition that

$$\sum_{\alpha=1}^{N} {}_m\delta_{\alpha,i} \leqq 1, \qquad i = 1, \cdots, L, \tag{11}$$

that is, that none of the sources reports more than one individual, $\hat{N}_m$ is a more efficient estimator than $\hat{N}_c$. The following inequality is derived in the appendix:

$$R < \frac{1}{N} \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} \leqq 1, \tag{12}$$

where R represents the ratio of the sampling variance of $\hat{N}_m$ to the sampling variance of $\hat{N}_c$ for a simple random sample design.

5. Multiplicity estimators for rare vital events

The prospects of increasing the efficiency of survey estimates of rare health conditions by means of multiplicity estimators, prompted the NCHS to consider applying the multiplicity survey to the problem of estimating rates of rare vital events. In an earlier section of this paper, the dual sample survey method of estimating vital rates was described. According to that method, it will be recalled, estimates of the numerators of vital rates are based on sample surveys linked to vital records and estimates of the population denominators are usually based on household sample surveys. Actually, there is some redundancy in this method because the household survey can produce estimates of the numerators as well as the denominators. However, vital statistics based on household sample surveys are subject to large sampling errors because vital events are relatively rare.

In the single time household survey of population change, vital events that occurred during a preceding reference period are reported by the sample households. In the conventional household survey, counting rules are adopted which assure that each vital event that occurred during the reference period is uniquely linked to one household. For example, the conventional rule for counting deaths

links the death that occurred during the reference period to the former dwelling unit of the decedent and the conventional rule for counting births links the surviving baby to its dwelling unit of residence.

Recently, the NCHS began to apply the principles of the design of multiplicity surveys to single time household sample surveys of population change. We have been exploring the feasibility and effectiveness of alternative multiplicity rules for linking vital events to households. One kind of multiplicity rule is based on consanguine relationships and it links the vital event to households containing its relatives. For example, the consanguine rules adopted in the survey might state that "births are reported by parents and grandparents" and "deaths are reported by the spouse, siblings, and children." Accordingly, births would be reported in the survey by households containing either the parents or grandparents and deaths by households containing either the surviving sib, spouse or child. The household reporting a vital event in the survey would also report as ancillary information the number of other households that would be eligible to report the same event in compliance with the multiplicity rule. Thus, the household reporting a death would also report the number of other households containing either the surviving spouse or a sib or a child of the decedent. The ancillary information would be used to determine the multiplicity of the reported event.

An experimental survey was conducted to investigate the effect of alternative consanguine rules on the reliability as well as the validity of birth and death statistics collected in single time household sample surveys of population change. Some preliminary findings of the experiment have been published [20] which compare the completeness of coverage of white deaths associated with different counting rules. The results indicate that coverage of white deaths is more complete in multiplicity surveys based on specified consanguine rules than on conventional counting rules.

6. Summary and conclusions

National health and vital statistics are collected by the NCHS in a family of sample surveys and registration systems. Within the framework of these data collection systems, it is frequently not feasible to produce reliable estimates of rarely occurring vital events and rarely prevailing health conditions. Consequently, special sample survey strategies have been developed for dealing with the problem. These strategies, which are described in this paper, would be applicable to epidemiological studies of pollutant effects.

To some extent these strategies represent the application to health surveys of well known methods for increasing the efficiency of sample surveys to estimate rare items. One method involves assembling sample frames that decrease the rarity of the item and that provide information about the listed units which can be used in the sample selection, estimation and data collection processes to improve the efficiency of the survey. This technique has been applied in surveys

to estimate rare events and rare health conditions. For the former, the files of vital records serve as the sampling frame and for the latter, lists of medical sources serve as the sampling frame.

The multiplicity survey described in this report is a relatively new type of sample survey strategy for improving the accuracy of estimates of rare attributes that is being investigated by the NCHS. The multiplicity survey places a premium on counting rules which link several enumeration sources to the same individual with the rare attribute. In contrast, counting rules of the conventional survey prescribe that each individual is uniquely linked to a single source. The estimator of the multiplicity survey has served as an unbiased estimator in surveys with unavoidable duplicate reporting [4] and in surveys where sampling frames contain duplicate listings [6]. Not until recently, however, has the estimator been applied in multiplicity surveys where duplicate reporting is incorporated into the survey as a deliberate strategy to improve the accuracy of the survey estimates.

◇ ◇ ◇ ◇ ◇

APPENDIX

A simple random sample of ℓ out of L enumeration sources is selected without replacement. Unbiased estimators of N, the number of individuals in the population with a specified attribute, are

$$\hat{N}_\theta = \frac{L}{\ell}\sum_{j=1}^{\ell} {}_\theta\lambda_j, \qquad \theta = c, m \tag{A.1}$$

where ${}_c\lambda_j$ and ${}_m\lambda_j$ are defined in the text by (2) and (6) respectively. The variance of $\hat{N}_\theta$ is

$$V(\hat{N}_\theta) = \frac{L^2}{\ell}\frac{L-L}{L-1}\left\{\frac{1}{\ell}\sum_{i=1}^{L} {}_\theta\lambda_i^2 - \left(\frac{N}{L}\right)^2\right\}. \tag{A.2}$$

If a conventional rule were selected in the survey

$$\sum_{i=1}^{L} {}_c\lambda_i^2 = \sum_{i=1}^{L}\left(\sum_{\alpha=1}^{N} {}_c\delta_{\alpha,i}\right)^2 = N + \sum_{i=1}^{L}\sum_{\alpha\neq\beta}^{N} {}_c\delta_{\alpha,i}\,{}_c\delta_{\beta,i} \tag{A.3}$$

where ${}_c\delta_{\alpha,i}$ is defined in the text by (3). It follows that

$$V(\hat{N}_c) = \frac{L^2}{\ell}\frac{L-\ell}{L-1}\left\{P(1-P) + \sum_{i=1}^{L}\sum_{\alpha\neq\beta}^{N} {}_c\delta_{\alpha,i}\,{}_c\delta_{\beta,i}\right\} \tag{A.4}$$

where $P = N/L$ represents the average number of individuals reported per source.

If a multiplicity rule were selected such that no source reported more than one individual with the attribute, the following condition would be satisfied:

$$\text{(A.5)} \qquad \sum_{i=1}^{L} \sum_{\alpha \neq \beta}^{N} {}_m\delta_{\alpha,i}\, {}_m\delta_{\beta,i} = 0$$

where ${}_m\delta_{\alpha,i}$ is defined in the text by (7). It follows that

$$\text{(A.6)} \qquad \sum_{i=1}^{L} {}_m\lambda_i^2 = \sum_{i=1}^{L} \left(\sum_{\alpha=1}^{N} \frac{{}_m\delta_{\alpha,i}}{s_\alpha} \right)^2 = \sum_{\alpha=1}^{N} \sum_{i=1}^{L} \left(\frac{{}_m\delta_{\alpha,i}}{s_\alpha} \right)^2 = \sum_{\alpha=1}^{N} \frac{1}{s_\alpha}$$

where $s_\alpha \geqq 1$ is defined in the text by (8). Consequently, we write

$$\text{(A.7)} \qquad V(\hat{N}_m) = \frac{L^2}{\ell} \frac{L-\ell}{L-1} \left\{ \frac{1}{L} \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} - \left(\frac{N}{L} \right)^2 \right\}$$

$$= \frac{L^2}{\ell} \frac{L-\ell}{L-1} P \left\{ \frac{1}{N} \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} - P \right\}.$$

Assuming that no source reports more than one individual in the multiplicity survey,

$$\text{(A.8)} \qquad R = \frac{V(\hat{N}_m)}{V(\hat{N}_c)} = \frac{P \left\{ \frac{1}{N} \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} - P \right\}}{P(1-P) + \sum_{i=1}^{L} \sum_{\alpha \neq \beta}^{N} {}_c\delta_{\alpha,i}\, {}_c\delta_{\beta,i}}$$

$$\leqq \frac{\frac{1}{N} \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} - P}{1-P} < \frac{1}{N} \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} \leqq 1.$$

REFERENCES

[1] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, New York, Wiley, 1966.

[2] L. A. Goodman, "Snowball sampling," *Ann. Math. Statist.*, Vol. 32 (1961), pp. 148–170.

[3] W. Haenszel, D. B. Loveland, and M. G. Sirken, "Lung cancer mortality as related to residence and smoking histories, I. white males," *J. Nat. Cancer Inst.*, Vol. 28 (1962), pp. 947–1001.

[4] M. H. Hansen, W. N. Hurwitz, and W. G. Madow, *Sample Survey Methods and Theory*, Vol. 1, New York, Wiley, 1953.

[5] Nan-Chang Hsieh, *Some Estimation Techniques for Utilizing Information from Elements Not in the Sample*, Survey Research Center, UCLA, 1970.

[6] L. Kish, *Survey Sampling*, New York, Wiley, 1965.

[7] National Center for Health Statistics, "Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates," *Vital and Health Statistics*, Ser. 2, No. 11 (1968), pp. 1–8.

[8] ———, "The 1968 revision of the standard certificates," *Vital and Health Statistics*, Ser. 4, No. 8 (1968), pp. 1–47.

[9] ———, "Methods and response characteristics, National Natality Survey United States 1963," *Vital and Health Statistics*, Ser. 22, No. 3 (1966), pp. 1–36.

[10] ———, "Medical X-ray visits and examinations during pregnancy, United States 1963," *Vital and Health Statistics*, Ser. 22, No. 5 (1968), pp. 1–41.

[11] ———, "Infant mortality rates by socioeconomic status," *Vital and Health Statistics*, Ser. 22, in print.

[12] ———, "Origin, program, and operation of the U.S. National Health Survey," *Vital and Health Statistics*, Ser. 1, No. 1 (1963), pp. 1-41.
[13] M. G. SIRKEN, "Sampling survey program of the National Vital Statistics Division," *Proceedings of the 9th National Meeting of the Public Health Conference on Records and Statistics*, 1962, pp. 39-41.
[14] ———, "Household surveys with multiplicity," *J. Amer. Statist. Assoc.*, Vol. 65 (1970), pp. 257-266.
[15] ———, "Stratified sample surveys with multiplicity," *J. Amer. Statist. Assoc.*, Vol. 67 (1972), in print.
[16] M. G. SIRKEN, J. W. PIFER, and M. L. BROWN, "Survey procedures for supplementing mortality statistics," *Amer. J. Pub. Health*, Vol. 50 (1960), pp. 1753-1764.
[17] M. G. SIRKEN and M. L. BROWN, "Quality of data elicited by successive mailings in mail surveys," *Proc. Soc. Statist. Sec. Amer. Statist. Assoc.* (1962), pp. 118-125.
[18] M. G. SIRKEN and H. L. DUNN, "Expanding and improving vital statistics," *Pub. Health Rep.* (1958), pp. 537-540.
[19] M. G. SIRKEN, J. W. PIFER, and M. L. BROWN, *Design of Surveys Linked to Death Records*, U.S. Department of Health, Education, and Welfare, Public Health Service, 1962.
[20] M. G. SIRKEN and P. N. ROYSTON, "Reasons deaths are missed in household surveys of population change," *Proc. Soc. Statist. Sec. Amer. Statist. Assoc.* (1970), pp. 361-364.

Discussion

Question: A. C. Hexter, California Department of Public Health

If a source which is supposed to report an event does not (if, for example, physician failed to report one of his patients), would that not lead to bias?

Reply: M. G. Sirken

Yes.

Question: S. Raman, Division of Biostatistics, University of California, Berkeley

Does your scheme of multiple reporting include the case where the same source reports more than once about the same patient?

Reply: M. G. Sirken

Yes.