NATIONAL CENTER For HEALTH STATISTICS | Series 2 Number 11

VITAL and HEALTH STATISTICS
DATA EVALUATION AND METHODS RESEARCH

Design of Sample Surveys to Estimate the Prevalence of Rare Diseases:

Three Unbiased Estimates

Three formulas, varying in the extent that they utilize all the information collected by a survey based on a stratified sample of medical sources, are presented to estimate the number of diagnosed cases of a rare disease in the population.

Washington, D.C. October 1965

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE
John W. Gardner
Secretary

Public Health Service
Luther L. Terry
Surgeon General

Public Health Service Publication No. 1000-Series 2-No. 11

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C., 20402 - Price 15 cents

NATIONAL CENTER FOR HEALTH STATISTICS

Public Health Service Publication No. 1000-Series 2-No. 11
Library of Congress Catalog Card Number 65-61845

CONTENTS

IN THIS REPORT a stratified random sample design is proposed for a survey of medical sources to estimate the prevalence of diagnosed cases of a rare disease in the population. The medical sources are stratified by criteria, such as specialty of physician and service and size of hospital, which are presumed to be related to the probability of the source treating patients with the particular disease. Deriving an unbiased estimate of the number of diagnosed cases based on patients reported by a sample of medical sources, however, presents a special problem since it is not uncommon for patients with rare diseases to have been treated by more than one medical source. Three formulas are presented for deriving unbiased estimates under these circumstances. The formulas differ in the extent that they utilize information collected in the sample survey about multiple reporting of patients.

DESIGN OF SAMPLE SURVEYS TO ESTIMATE THE PREVALENCE OF RARE DISEASES:

THREE UNBIASED ESTIMATES

Z. W. Birnbaum, Ph.D. *Department of Mathematics, University of Washington*
Monroe G. Sirken, Ph.D., *Division of Health Records Statistics*

1. INTRODUCTION

The National Center for Health Statistics has been concerned with the problem of developing survey procedures for estimating the occurrence of diagnosed cases of rare diseases in the population and, during the past few years, has conducted several sample surveys to obtain information on diagnosed cases of cystic fibrosis.[1] [2] The incidence of this genetic disease in the population had been variously estimated as occurring in from 1 in 3,700 live births[3] to 1 in 600 live births.[4] On the basis of a survey conducted in three New England States,[2] it was estimated that 651 cystic fibrosis patients had received medical care during the 8-year period 1952-59, indicating an incidence of 1 diagnosed case per 2,300 live births and a high mortality rate during the early years of life. Although a relatively rare disease, cystic fibrosis has been increasingly recognized as one of the major childhood diseases and as representing an important health problem since it requires intensive medical care and expensive treatment and nearly always results in premature death.

Historically, the design of sample surveys to estimate the occurrence of diagnosed cases of rare diseases in the population has presented two major difficulties: (a) large sampling errors usually associated with survey estimates of rarely occurring events and (b) potentially large nonsampling errors associated with diagnostic information reported in surveys. To minimize these difficulties, a stratified random sample design of medical sources has been used in cystic fibrosis surveys conducted by the Center. Medical sources, the primary sampling units in these surveys, were stratified according to the likelihood of their treating cases of this disease of children. Thus, pediatricians and hospitals with a pediatric residency were selected with certainty, while in other strata varying fractions of physicians and hospitals were selected for the sample. In order to control errors of diagnosis, medical sources were asked to provide the results of diagnostic tests for each reported cystic fibrosis case. This information was subsequently reviewed by clinicians to substantiate or contraindicate the diagnosis. Where the reported diagnostic information was incomplete, the medical source was frequently asked to provide the missing information if it was available.

Deriving an estimate of the number of diagnosed cases of a rare disease from the sample survey of medical sources presents a unique statistical problem because it is not uncommon for the patient to have been treated and hence to be reported by several medical sources. For example, the results of a cystic fibrosis survey in three New England States conducted by the Center indicated that each of the 651 patients had been treated, on the average, by 1.6 sources. The percentage distribution of these patients by the number of medical sources who reported them was as follows:

	Percentage distribution
Total patients----------	100.0
Patients reported by 1 source------	53.6
Patients reported by 2 sources------	32.5
Patients reported by 3 sources-----	10.8
Patients reported by 4 or 5 sources--	3.1

Under these circumstances, it appears to be necessary to obtain information about the "multiplicity" of medical sources that treated the patients reported in the sample survey in order to derive unbiased estimates of the number of diagnosed cases.

Recent studies conducted by the Center suggest the feasibility of a particular survey procedure to obtain this kind of information for cases of cystic fibrosis, and it appears very likely that this same procedure could be readily adapted to surveys for other rare diseases.

According to this survey procedure, each medical source in the sample survey reports separately each of the cystic fibrosis patients that he has treated. For each patient reported, he identifies other medical sources that have also treated the patient. Subsequently, the coverage of the survey is expanded to include those medical sources not originally selected in the sample (nonsample medical sources) that have been identified as having treated patients reported by the medical sources in the original sample. The purpose of the survey of nonsample medical sources is to verify that they in fact did treat and do report the patient(s) in question, and if they do report the patients, to have them identify other medical sources that also treated these patients.

In this manner, the survey presumably is able to identify all of the medical sources, whether or not they are included in the original sample, that treated patients reported by one or more of the original sample medical sources. The success of this survey procedure hinges ultimately on the extent to which medical sources that have treated the same patients for a specified disease know of each other and report one another in the survey. In view of the seriousness and infrequent occurrence of the kind of diseases being considered here, this assumption seems reasonable. More developmental work is needed, however, to test the survey method for completeness in identifying the medical sources treating the same patient and to determine the size of the bias, due to incomplete identification of medical sources, in the estimate of the number of patients with the disease.

Three unbiased estimates of the number of diagnosed cases of a rare disease in the population based on a stratified sample of medical sources are presented in this report. Each of them makes use of information on "multiplicity" of sources reporting a patient obtained in the manner described above from cystic fibrosis surveys of medical sources. The first estimate (see formula 4.6) utilizes the least information on "multiplicity;" it requires only the total number of medical sources that treated each patient reported by a medical source in the original sample. The second estimate (see formulas 5.4 and 5.5) requires information on the number of medical sources within each stratum that treated the same patient reported by a sample source. The third estimate (see formulas 6.7 and 6.8) appears to utilize the most detailed "multiplicity" information. The second and third estimates require the matching of patients reported by sample sources to assure that each patient is counted only once. For the first estimate, each patient is counted as many times as he is reported by sample sources.

This is a report of some preliminary results from research work now in progress. Besides the estimates presented in this report, still other estimates may exist which use in more detail the "multiplicity" information collected in the survey than do the estimates presented here. The variances of the three proposed estimates are not compared although the variances of the first and second estimates are derived in this report. These results are preliminary to work on the problem of optimum allocation of sample size among the strata. Existing statistical theory[5,6] on optimum allocation in stratified random sampling does not consider the possibility of overlapping of elements (patients) among the primary sample units (medical sources) either in the same stratum or in different strata.

2. DESCRIPTION OF THE POPULATION

The sampling units are "medical sources" such as physicians in individual practice, clinics, or hospitals. The population of all such medical sources consists of strata, for example, large hospitals, small hospitals, pediatricians, and other physicians. To describe this population we use the following notations:

$$\Pi = \text{population of all medical sources,}$$

$$\xi_r = r\text{-th stratum, } r = 1,2,\ldots,R, \quad (2.1)$$

$$S_{r,i} = \text{the } i\text{-th source of the stratum } \xi_r \quad i = 1,2,\ldots,M_r.$$

The population Π is therefore the union of the R strata $\xi_1, \xi_2, \ldots, \xi_R$; the $r-$th stratum ξ_r consists of the M_r sources $S_{r,1}, S_{r,2}, \ldots, S_{r,M_r}$; and Π consists of all the sources $S_{1,1}, S_{1,2}, \ldots, S_{1,M_1}, S_{2,1}, S_{2,2}, \ldots, S_{2,M_2}, \ldots, S_{R,1}, S_{R,2}, \ldots, S_{R,M_R}$.

The number of sources in ξ_r is M_r, and the total number of sources in Π is

$$M = \sum_{r=1}^{R} M_r . \tag{2.2}$$

The different individual patients are denoted by $I_1, I_2, \ldots, I_N$

where

$$N = \text{number of different patients} \tag{2.3}$$

is the quantity we wish to estimate. When a patient I_α appears in the records of a medical source $S_{r,i}$, we shall say that he is "contained" in that source. Since each patient may be contained in several sources, we introduce the indicator variables

$$\mu_{\alpha,r,i} = \begin{cases} 1 \text{ when } I_\alpha \text{ is contained in } S_{r,i} \\ 0 \text{ when } I_\alpha \text{ is not contained in } S_{r,i}. \end{cases} \tag{2.4}$$

These indicator variables are defined for $\alpha = 1,2, \ldots, N$; $r = 1,2, \ldots, R$; $i = 1,2, \ldots, M_r$, and with their aid we define the auxiliary quantities listed below together with their intuitive meanings which are easy to verify:

$$\sum_{\alpha=1}^{N} \mu_{\alpha,r,i} = N_{r,i} = \text{number of different patients in } S_{r,i}, \tag{2.5.1}$$

$$\sum_{i=1}^{M_r} \mu_{\alpha,r,i} = s_{\alpha,r} = \text{number of sources in } \xi_r \text{ containing } I_\alpha . \tag{2.5.2}$$

$$\sum_{r=1}^{R} \sum_{i=1}^{M_r} \mu_{\alpha,r,i} = \sum_{r=1}^{R} s_{\alpha,r} = s_\alpha = \text{number of sources in } \Pi \text{ containing } I_\alpha . \tag{2.5.3}$$

It is easily seen that

$$N = \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} \sum_{r=1}^{R} s_{\alpha,r} = \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} \sum_{r=1}^{R} \sum_{i=1}^{M_r} \mu_{\alpha,r,i} . \tag{2.6}$$

The totality of the $\mu_{\alpha,r,i}$, for $\alpha = 1, \ldots, N$; $r = 1, \ldots, R$; $i = 1, \ldots, M_r$, represents a complete description of the way in which every patient occurs or does not occur in the records of each of the sources. Clearly in view of (2.6) the knowledge of all the $\mu_{\alpha,r,i}$ would imply the knowledge of N.

3. THE SAMPLING SURVEY

The sampling design used consists of a one-stage stratified sample from Π, and the survey consists of the following steps.

a. From the M_r sources of the stratum ξ_r a number $m_r (\leq M_r)$ is drawn at random without replacements for $r = 1,2, \ldots, R$. The samples from different strata are independent. The total sample size is

$$m = \sum_{r=1}^{R} m_r .$$

b. If a source $S_{r,i}$ is included in the sample from ξ_r then all its records are reviewed, and the record for every patient I_α suffering from the disease under study and contained in $S_{r,i}$ is obtained.

c. This record of I_α contained in a source $S_{r,i}$ which occurs in the sample is used to determine all other sources which contain the same patient I_α, no matter whether these other sources occur in the sample or not.

The description of step c may be paraphrased by saying that when patient I_α is contained in a source $S_{r,i}$ which occurs in the sample, we not only learn that $\mu_{\alpha,r,i} = 1$, but we also find out from the record of I_α in $S_{r,i}$ the values of *all* $\mu_{\alpha,t,j}$ for this patient I_α and all sources $S_{t,j} \neq S_{r,i}$.

The information obtained in step c is the essential feature of the design, and by utilizing this information, more or less completely, we are able to construct the estimates for N described and discussed in the following sections.

4. AN UNBIASED ESTIMATE FOR N

According to (2.5.3) patient I_α belongs to s_α sources in Π, where s_α may be any of the integers $1,2, \ldots, M$. When $s_\alpha = s$, we shall say

that I_α has "multiplicity s" in Π, or that he is an "s-fold patient" in Π. We now define

$$n_{s,r,i} = \text{number of } s\text{-fold patients in } S_{r,i} \tag{4.1}$$

and

$$n_{r,i} = \sum_{s=1}^{M} \frac{1}{s}\, n_{s,r,i} . \tag{4.2}$$

One verifies easily that

$$N = \sum_{r=1}^{R} \sum_{i=1}^{M_r} n_{r,i} . \tag{4.3}$$

We futhermore introduce the random variables

$$U_{r,i} = \begin{cases} 1 \text{ when } S_{r,i} \text{ occurs in the sample} \\ 0 \text{ when } S_{r,i} \text{ does not occur in the sample} \end{cases} \tag{4.4}$$

with the probabilities

$$p_{r,i} = P_r\,\{U_{r,i} = 1\} = E\,\{U_{r,i}\} = \frac{m_r}{M_r} \tag{4.5}$$

for $r = 1, \ldots, R$; $i = 1, \ldots, M_r$. The probabilities (4.5) are determined by the sample design.

We now consider the statistic

$$\tilde{N} = \sum_{r=1}^{R} \sum_{i=1}^{M_r} \frac{n_{r,i}}{p_{r,i}}\, U_{r,i} . \tag{4.6}$$

While the right-hand side of (4.6) formally is a sum over all sources in Π, it can be computed by using only the information obtained from those sources which occur in the sample, since for all other sources one has $U_{r,i} = 0$ and the corresponding term in (4.6) vanishes. When $S_{r,i}$ occurs in the sample, we can determine for every patient contained in it his multiplicity, hence we can compute the $n_{s,r,i}$ and $n_{r,i}$, and thus the coefficient of $U_{r,i}$ in (4.6). In view of (4.5) and (4.3) one has immediately

$$E\,(\tilde{N}) = \sum_{r=1}^{R} \sum_{i=1}^{M_r} \frac{n_{r,i}}{p_{r,i}}\, E\,(U_{r,i}) = N \tag{4.7}$$

hence $\tilde{N}$ is an unbiased estimate for N.

Whenever source $S_{r,i}$ occurs in the sample, we can according to step c of our survey (Section 3) determine not only the multiplicity of every patient in $S_{r,i}$ —which is all that is needed to compute $n_{r,i}$, and hence $\tilde{N}$ —but we can also tell for every other source in Π whether or not it contains any one patient contained in $S_{r,i}$. This more detailed information derived from the survey is not utilized in the statistic $\tilde{N}$.

To compute the variance of $\tilde{N}$, we first note that $n_{r,i}$ is a variate ascribed to the source $S_{r,i}$, and so is $\frac{n_{r,i}}{p_{r,i}} = n_{r,i} \cdot \frac{M_r}{m_r}$. The statistic $\tilde{N}$ is therefore the sum of the values of the variate $n_{r,i} \cdot \frac{M_r}{m_r}$ obtained by taking a stratified sample without replacements from Π. Applying the classical theory of such samples one obtains

$$\operatorname{var}(\tilde{N}) = \sum_{r=1}^{R} \frac{M_r - m_r}{m_r(M_r - 1)} \left[M_r \sum_{i=1}^{M_r} (n_{r,i})^2 - \Big(\sum_{i=1}^{M_r} n_{r,i}\Big)^2 \right]. \tag{4.8}$$

For given total sample size $m = \sum_{r=1}^{R} m_r$ the right side in (4.8) is minimum when

$$m_r = m \cdot \frac{\sqrt{A_r}}{\sum_{s=1}^{R} \sqrt{A_s}}$$

where

$$A_r = \frac{M_r}{M_r - 1} \left[M_r \sum_{i=1}^{M_r} (n_{r,i})^2 - \Big(\sum_{i=1}^{M_r} n_{r,i}\Big)^2 \right]$$

for $r = 1, \ldots, R$.

5. ANOTHER UNBIASED ESTIMATE FOR N

We now consider for each patient I_α the two complementary events: (a) he is contained in at least one of the sources occurring in the sample, and (b) he is not contained in any of the sources of the sample. The following indicator variable tells which of these two events has occurred:

$$W_\alpha = \begin{cases} 1 \text{ when } I_\alpha \text{ is in a source included in the sample} \\ 0 \text{ otherwise.} \end{cases} \tag{5.1}$$

The random variable W_α has the probability distribution

$$Pr\ \{W_\alpha = 0\} = \prod_{r=1}^{R} \frac{\binom{M_r - s_{r,\alpha}}{m_r}}{\binom{M_r}{m_r}} = q_\alpha ,$$

$$Pr\ \{W_\alpha = 1\} = 1 - q_\alpha = p_\alpha . \tag{5.2}$$

A linear estimate of N in terms of the W_α

$$\sum_{\alpha=1}^{N} \gamma_\alpha W_\alpha \tag{5.3}$$

is unbiased if and only if

$$\sum_{\alpha=1}^{N} \gamma_\alpha\ E\ (W_\alpha) = \sum_{\alpha=1}^{N} \gamma_\alpha p_\alpha = N$$

which is an identity in the p_α if and only if $\gamma_\alpha = \frac{1}{p_\alpha}$ for $\alpha = 1, \ldots, N$. We thus arrive at a unique unbiased linear estimate for N of the form (5.3)

$$\overset{v}{N} = \sum_{\alpha=1}^{N} \frac{1}{p_\alpha}\ W_\alpha \tag{5.4}$$

where

$$p_\alpha = 1 - q_\alpha = 1 - \prod_{r=1}^{R} \frac{\binom{M_r - s_{\alpha,r}}{m_r}}{\binom{M_r}{m_r}} . \tag{5.5}$$

As in (4.6), the right-hand side of (5.4) can again be computed from the information provided in our survey. It utilizes, however, a different aspect of this information, since in computing the coefficients $\frac{1}{p_\alpha}$ in (5.4) according to formula (5.5) one needs, for each I_α contained in a source occurring in the sample, all the quantities

$$s_{\alpha,r} = \text{multiplicity of } I_\alpha \text{ in } \xi_r$$

for $r = 1, \ldots, R$, while for computing the coefficients in (4.6) one needed for every source $S_{r,i}$ occurring in the sample all quantities

$$n_{s,r,i} = \text{number of s-fold patients in } S_{r,i}$$

for $s = 1, 2, \ldots, M$.

The variance of $\overset{v}{N}$ can be computed as follows:

$$\operatorname{var}(\overset{v}{N}) = \sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} \operatorname{cov}\left(\frac{1}{p_\alpha} W_\alpha, \frac{1}{p_\beta} W_\beta\right) =$$

$$= \sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} \frac{1}{p_\alpha} \cdot \frac{1}{p_\beta} \operatorname{cov}\ (W_\alpha, W_\beta)$$

where, for $\alpha \neq \beta$

$$\operatorname{cov}\ (W_\alpha, W_\beta) = E\ (W_\alpha W_\beta) - E\ (W_\alpha)\ E\ (W_\beta) =$$

$$= Pr\ \{W_\alpha = W_\beta = 1\} - p_\alpha p_\beta .$$

To write $Pr\{W_\alpha = W_\beta = 1\}$ one needs the auxiliary quantities

$$s_{\alpha,\beta,r} = \sum_{i=1}^{M_r} \mu_{\alpha,r,i} \cdot \mu_{\beta,r,i} = \tag{5.6}$$

= number of sources containing both I_α and I_β.

One has

$$p_{\alpha,\beta} = Pr\ \{W_\alpha = W_\beta = 1\} =$$

$$= 1 - Pr\ \{W_\alpha = 0\} - Pr\ \{W_\beta = 0\} +$$

$$+ Pr\ \{W_\alpha = W_\beta = 0\} =$$

$$= 1 - q_\alpha - q_\beta + \tag{5.7}$$

$$+ \prod_{r=1}^{R} \frac{\binom{M_r - s_{\alpha,r} - s_{\beta,r} + s_{\alpha,\beta,r}}{m_r}}{\binom{M_r}{m_r}}$$

and

$$\operatorname{cov}\ (W_\alpha, W_\beta) = p_{\alpha,\beta} - p_\alpha p_\beta . \tag{5.8}$$

For $\alpha=\beta$ obviously

$$\text{cov}\ (W_\alpha, W_\alpha) = \text{var}\ (W_\alpha) = p_\alpha q_\alpha$$

and with the conventions

$$s_{\alpha,\alpha,r} = s_{\alpha,r}$$

$$p_{\alpha,\alpha} = p_\alpha$$

one obtains

$$\text{var}\ (\overset{\vee}{N}) = \sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} \frac{p_{\alpha,\beta} - p_\alpha p_\beta}{p_\alpha p_\beta} . \qquad (5.9)$$

6. AN UNBIASED ESTIMATE BASED ON STRATA WITH PRIORITIES

All the sources $S_{r,i}$ may be ordered in a sequence by ascribing them priorities in some arbitrary manner. We choose the following convention.

Priority Rule

Source $S_{q,i}$ has higher priority than source $S_{r,j}$ when either $q < r$ or $q = r$ and $i < j$.

Since every patient I_α may be contained in several sources, and we wish to account for this in our estimate of N, we now decide to count him not more than once (and later on to attach a weight to this count) according to the following counting rule.

If patient I_α is contained in some sources occurring in the sample, he is counted only in that one among them which has the highest priority.

We now define the indicator variables

$$V_{\alpha,r,i} = \begin{cases} 1 \text{ when } I_\alpha \text{ is counted in } S_{r,i} \\ 0 \text{ otherwise} \end{cases} \qquad (6.1)$$

and consider linear estimates of the form

$$\hat{N} = \sum_{\alpha=i}^{N} \sum_{r=1}^{R} \sum_{i=1}^{M_r} \gamma_{\alpha,r,i}\ V_{\alpha,r,i} . \qquad (6.2)$$

To compute the probabilities of the random variables (6.1) we note that $V_{\alpha,r,i} = 1$ if and only if I_α is contained in $S_{r,i}$, and $S_{r,i}$ occurs in the sample, and $S_{r,i}$ has highest priority among the sources occurring in the sample which contain I_α. We have therefore

$$Pr\ \{V_{\alpha,r,i} = 1\} = E\ \{V_{\alpha,r,i}\} =$$

$$= \mu_{\alpha,r,i} \cdot Pr\ \{S_{r,i} \text{ is in the sample}\}.$$

$$\cdot Pr\ \{S_{r,i} \text{ has highest priority among sources in the sample which contain } I_\alpha, \text{ given that } S_{r,i} \text{ is in the sample}\}.$$

Denoting the last of the three factors by $\pi_{\alpha,r,i}$, we write this in the form

$$Pr\ \{V_{\alpha,r,i} = 1\} = \mu_{\alpha,r,i} \cdot \frac{m_r}{M_r} \cdot \pi_{\alpha,r,i} . \qquad (6.3)$$

Of the three factors in (6.3) neither $\frac{m_r}{M_r}$ nor $\pi_{\alpha,r,i}$ depend on whether I_α is contained in $S_{r,i}$.

To evaluate $\pi_{\alpha,r,i}$ we introduce the auxiliary quantities

$$\sum_{j=1}^{i-1} \mu_{\alpha,r,j} = s^{(i)}_{\alpha,r} = \text{number of sources in } \xi_r \text{ containing } I_\alpha \text{ and having higher priority than } S_{r,i} \qquad (6.4)$$

for $\alpha = 1, \ldots, N$; $r = 1, \ldots, R$; $i = 1, \ldots, M_r$.

One verifies that

$$\pi_{\alpha,1,i} = \frac{\binom{M_1 - 1 - s^{(i)}_{\alpha,1}}{m_1 - 1}}{\binom{M_1 - 1}{m_1 - 1}} \qquad (6.5)$$

and

$$\pi_{\alpha,r,i} = \frac{\binom{M_r - 1 - s^{(i)}_{\alpha,r}}{m_r - 1}}{\binom{M_r - 1}{m_r - 1}} \prod_{g=1}^{r-1} \frac{\binom{M_g - s_{\alpha,g}}{m_g}}{\binom{M_g}{m_g}} \qquad (6.6)$$

for $r \geq 2$.

In view of (2.6), for (6.2) to be an unbiased estimate for N it is necessary and sufficient that

$$\sum_{\alpha=1}^{N} \sum_{r=1}^{R} \sum_{i=1}^{M_r} \gamma_{\alpha,r,i}\ \mu_{\alpha,r,i} \cdot \frac{m_r}{M_r} \cdot \pi_{\alpha,r,i} = N =$$

$$= \sum_{\alpha=1}^{N} \sum_{r=1}^{R} \sum_{i=1}^{M_r} \cdot \frac{1}{s_\alpha} \cdot \mu_{\alpha,r,i}$$

identically in $\mu_{\alpha,r,i}$, hence

$$\gamma_{\alpha,r,i} \cdot \frac{m_r}{M_r} \cdot \pi_{\alpha,r,i} = \frac{1}{s_\alpha}$$

and the unique unbiased estimate of the form (6.2) is

$$\hat{N} = \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} \sum_{r=1}^{R} \frac{M_r}{m_r} \sum_{i=1}^{M_r} \frac{1}{\pi_{\alpha,r,i}} V_{\alpha,r,i}. \qquad (6.7)$$

The following comments on this estimate may be in order.

a. For every α, at most one term of the double sum

$$\sum_{r=1}^{R} \frac{m_r}{M_r} \sum_{i=1}^{M_r} \frac{1}{\pi_{\alpha,r,i}} V_{\alpha,r,i}$$

can be 1, and all others must vanish, since I_α can be counted in at most one source. One can therefore write (6.7) in the form

$$\hat{N} = \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} \cdot \frac{M_\rho}{m_\rho} \cdot \frac{1}{\pi_{\alpha,\rho,\iota}} \cdot V_{\alpha,\rho,\iota} \qquad (6.8)$$

where (ρ,ι) are the subscripts of that source in which I_α is counted.

b. As in the estimates $\tilde{N}$ and $\check{N}$, the coefficients in (6.7) or (6.8) are themselves statistics, i.e., they can be computed from the information provided in the survey. However, to compute $\pi_{\alpha,\rho,\iota}$ for given α one actually needs to know a large number of the $\mu_{\alpha,r,i}$; hence the information obtained in the survey appears to be utilized here in considerable detail.

The variance of $\hat{N}$ can be written out by using auxiliary quantities similar to (6.4) but referring to pairs of patients I_α, I_β. The resulting expressions are unwieldy and we have not succeeded in putting them in a form which would lend itself to some intuitive interpretation.

7. SOME GENERAL OBSERVATIONS

In the sampling design of section 3, it is possible that for some strata of medical sources one decides to carry out a complete census, i.e., to set the sample size equal to the number of sources. If this is done, one may lump all such strata into one stratum ξ_o and call it the certainty stratum. This convention would imply

$$m_o \doteq M_o$$
$$m_r < M_r \text{ for } r = 1,2,\ldots,R.$$

It would also assure a complete enumeration of patients who belong to sources in ξ_o, and any effect of random sampling would be limited to those patients who do not belong to any sources in ξ_o. One can, therefore, assume without loss of generality that the developments of the preceding sections were made for designs which did not provide for a certainty stratum.

We do not know at this time how the variances of the three estimates proposed in sections 4, 5, and 6 compare. In fact, we do not know whether there are other estimates which would utilize the information obtained in the survey described in Section 3 in more detail than the estimates proposed here. Similarly, the problem of optimal allocation of sample sizes for the estimates $\check{N}$ and $\hat{N}$ remains open.

REFERENCES

[1]Sirken, M. G., Crane, M. M., Brown, M. L., and Kramm, E. R.: A national hospital survey of cystic fibrosis. *Pub. Health Rep.* 74:764-770, Sept. 1959.

[2]Kramm, E. R., Crane, M. M., Sirken, M. G., and Brown, M. L.: A cystic fibrosis pilot survey in three New England States. *Am. J. Pub. Health* 52:2041-2057, Dec. 1962.

[3]Steinberg, A. G., and Brown, D. C.: On the incidence of cystic fibrosis of the pancreas. *Am. J. Human Genet.* 12(4): 416-424, Dec. 1960.

[4]Anderson, D. H., and Hodges, R. G.: Celiac syndrome, V, Genetics of cystic fibrosis of the pancreas with a consideration of etiology. *Am. J. Dis. Child.* 72(1):62-80, July 1946.

[5]Hansen, M. H., Hurwitz, W. N., and Madow, W. G.: *Sample Survey Methods and Theory,* Volumes I and II. New York. John Wiley and Sons, Inc., 1963.

[6]Cochran, W. G.: *Sampling Techniques.* New York. John Wiley and Sons, Inc., 1953.

———— ooo ————

★ U.S. GOVERNMENT PRINTING OFFICE : 1965 O—786-114

OUTLINE OF REPORT SERIES FOR VITAL AND HEALTH STATISTICS

Public Health Service Publication No. 1000

Series 1. *Programs and collection procedures.*—Reports which describe the general programs of the National Center for Health Statistics and its offices and divisions, data collection methods used, definitions, and other material necessary for understanding the data.

Reports number 1-4

Series 2. *Data evaluation and methods research.*—Studies of new statistical methodology including: experimental tests of new survey methods, studies of vital statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, contributions to statistical theory.

Reports number 1-11

Series 3. *Analytical studies.*—Reports presenting analytical or interpretive studies based on vital and health statistics, carrying the analysis further than the expository types of reports in the other series.

Reports number 1-4

Series 4. *Documents and committee reports.*—Final reports of major committees concerned with vital and health statistics, and documents such as recommended model vital registration laws and revised birth and death certificates.

Reports number 1 and 2

Series 10. *Data From the Health Interview Survey.*—Statistics on illness, accidental injuries, disability, use of hospital, medical, dental, and other services, and other health-related topics, based on data collected in a continuing national household interview survey.

Reports number 1-22

Series 11. *Data From the Health Examination Survey.*—Statistics based on the direct examination, testing, and measurement of national samples of the population, including the medically defined prevalence of specific diseases, and distributions of the population with respect to various physical and physiological measurements.

Reports number 1-10

Series 12. *Data From the Health Records Survey.*—Statistics from records of hospital discharges and statistics relating to the health characteristics of persons in institutions, and on hospital, medical, nursing, and personal care received, based on national samples of establishments providing these services and samples of the residents or patients.

Reports number 1 and 2

Series 20. *Data on mortality.*—Various statistics on mortality other than as included in annual or monthly reports—special analyses by cause of death, age, and other demographic variables, also geographic and time series analyses.

No reports to date

Series 21. *Data on natality, marriage, and divorce.*—Various statistics on natality, marriage, and divorce other than as included in annual or monthly reports—special analyses by demographic variables, also geographic and time series analyses, studies of fertility.

Reports number 1-6

Series 22. *Data From the National Natality and Mortality Surveys.*—Statistics on characteristics of births and deaths not available from the vital records, based on sample surveys stemming from these records, including such topics as mortality by socioeconomic class, medical experience in the last year of life, characteristics of pregnancy, etc.

Reports number 1

For a list of titles of reports published in these series, write to: National Center for Health Statistics
U.S. Public Health Service
Washington, D.C. 20201