DATA EVALUATION AND METHODS RESEARCH

Series 2
Number 38

Estimate and Sampling Variance

in the Health Interview Survey

A method for computing variances of estimates derived from the Health Interview Survey.

DHEW Publication No. (HRA) 74-1288

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service

Health Resources Administration
National Center for Health Statistics
Rockville, Md. March 1974

Vital and Health Statistics Series 2-No. 38
First issued in the Public Health Service Publication Series No. 1000
June 1970

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price 65 cents

NATIONAL CENTER FOR HEALTH STATISTICS

COOPERATION OF THE BUREAU OF THE CENSUS

In accordance with specifications established by the National Health Survey, the Bureau of the Census, under a contractual agreement, participated in the design and selection of the sample, and carried out the first stage of the field interviewing and certain parts of the statistical processing.

Vital and Health Statistics-Series 2-No. 38

DHEW Publication No. (HRA) 74-1288
Library of Congress Catalog Card Number 76-605819

CONTENTS

THIS REPORT PRESENTS a method for computing variances of estimates derived from the Health Interview Survey (HIS). In addition it gives a brief description of the HIS sample design, a detailed account of how the estimator for health statistics is constructed, and an iterative method for fitting a curve to a set of variance estimates to obtain approximations that can be used for several different types of estimates derived from survey data.

The variance estimator is an adaptation of one initially developed by Professor Nathan Keyfitz. In applying the method to the Health Interview Survey, difficulties arise because certain fundamental assumptions are not met. The report tells how these problems were resolved, develops the theory which applies to the complex HIS estimator, and discusses some aspects for programming the procedure for the computer.

ESTIMATION AND SAMPLING VARIANCE IN THE HEALTH INTERVIEW SURVEY

Judy A. Bean, *Office of Statistical Methods*

INTRODUCTION

Purpose

The purpose of this report is to present a method for estimating variances for the Health Interview Survey. The report includes a discussion of the estimation process and of the procedures for implementing the variance estimating technique.

Complex estimation techniques are required for the Health Interview Survey, which is a highly stratified multistage probability sample of persons. The variance formulas must reflect features of the sample design, such as unequal sampling fractions, and components of the estimation procedure, such as ratio adjustment, nonresponse adjustment, and post-stratification. One technique proposed and used in this report is derived from the variance estimation method given by Nathan Keyfitz in an article[1] and is hereafter referred to as the Keyfitz method. This method is based on the premise that the variance of a sum of two values of a random variable is the expected value of the square of the difference between them. In applying the method to the particular case of the Health Interview Survey, difficulties arise because certain fundamental assumptions of the Keyfitz method are not met. The resolutions of these problems are described, and aspects of programming the procedure for the computer are discussed.

The application of the Keyfitz variance estimation method introduced here could be employed for other surveys similar to the Health Interview Survey.

Background

The program of the National Center for Health Statistics includes a variety of data collection techniques designed to assemble information on the health of the population. One facet of this program is the Health Interview Survey, a continuous sampling of the civilian, noninstitutional population to obtain information on illness, disability, and other health-related items.

The sampling design of the Health Interview Survey, which is conducted with the cooperation of the U.S. Bureau of the Census, produces a complex multipurpose survey. While the structure of the survey has developed since its inception in 1957, a basic design has persisted, with major modifications in 1959 and 1963. Originally a sample of 372 primary sampling units (PSU's), which consist of a county or a small group of contiguous counties collectively covering the 50 States and the District of Columbia, was selected from a universe of 1,900 units. The PSU's were divided geographically into units called segments, each containing an expected six households. Then a sample of approximately 6,000 segments yield-

ing 36,000 households was chosen from the sample PSU's. An interviewer's assignment for a work week usually consisted of two such segments.

In 1959 the number of primary sampling units was increased to 503 and the number of households to be interviewed during a year to 38,000. The average size of an assignment for an interviewer was increased from 12 to 13.5 households.

In 1963, when population data from the 1960 census became available, several changes were made in order to increase the efficiency of the sample design. The structure of segments and assignments was modified in three important respects: (1) segment size was changed from an expected six households to an expected nine households, (2) the nine households were alternate ones in a cluster of about 18 neighboring households, whereas earlier the six had been a compact cluster of adjacent households, and (3) assignments for a given week consisted of paired neighboring segments, while previously an assignment attempted to pair unlike segments (heterogeneity is obtained by assigning an interviewer different types of assignments in successive weeks). These changes resulted in an increase from 13.5 to 16 households in an average assignment.

Also in 1963 the manner of selecting specific segments was changed for about two-thirds of the total sample from area sampling to list sampling, using 1960 census registers as the list frame. Most of the remaining third of the sample continued as an area sample; in some sectors the sampling of building permits provided for newly constructed places.

The changes described above, together with benefits from joint designing with the Current Population Survey,[2] made it possible in 1963 to reduce the sample size to 357 PSU's and at the same time to increase the number of sample households during a year from 38,000 to 42,000 (in approximately 4,700 segments).

The sample design of the Health Interview Survey, the development of which has been summarized above, permits both continuous measurement of characteristics of high prevalence in the population and, through consolidation of samples, more detailed analysis of less common characteristics and smaller categories of health-related items. The continuous collection has administrative and operational advantages as well as technical assets since it permits field work to be handled with an experienced, stable staff. Descriptive material on data collection, field procedures, and questionnaire development in the Health Interview Survey can be found in an early report.[3]

Another publication issued during the early years of the interview survey, "The Statistical Design of the Health Household-Interview Survey,"[4] presents very detailed information on the structure, design, and procedures of the survey. This report also includes illustrative materials on estimation, measurement errors, stratification of sampling units, allocation of data, and selection of sampling units.

The present report deals with the technical aspects of the procedures used in deriving weighted estimates from the data collected in the sample and in measuring the reliability of the estimates produced.

ESTIMATION

Estimating Equation

Since the design of the Health Interview Survey is a complex multistage probability sample, it becomes necessary to use extensive procedures in the derivation of estimates. Four basic operations are involved.

Inflation by the reciprocal of the probability of selection.—This probability of selection is the product of the probabilities of selection from each step of selection in the design: PSU, segment, and household.

Nonresponse adjustment.—The estimates are inflated by a multiplication factor which has as its numerator the number of sample house-

holds in a given segment and as its denominator the number of households interviewed in that segment.

First-stage ratio adjustment.—Sampling theory indicates that the use of auxiliary information which is highly correlated with the variables being estimated improves the reliability of the estimates. To reduce the variability between PSU's within a region, the estimates are ratio adjusted to 1960 population within six color-residence classes.

Post-stratification by age-sex-color.—Here the estimates are ratio adjusted within each cell to an independent estimate of the population of the cell for the survey period. These independent estimates are prepared by the Census Bureau. Both the first-stage and post-stratified ratio adjustments take the form of multiplication factors applied to the weight of each elementary unit (person, household, condition, and hospitalization).

For illustration purposes the estimating equation for a nonresponse adjusted estimate, for a nonresponse first-stage ratio adjusted estimate, and for a nonresponse first-stage post-stratified estimate will be given.

1. Nonresponse adjusted estimate

$$\hat{x} = \sum_a \sum_c \sum_i \sum_h \sum_k w_{ih}\, x_{ihkca}\, \frac{n_{ik}}{n'_{ik}}$$

$$= \sum_a \sum_c \sum_i \sum_h \sum_k w'_{ih}\, x_{ihkca}$$

where

$\hat{x}$ = the nonresponse adjusted estimate of the x-health characteristic

w_{ih} = the weight of the hth person in the ith PSU; reciprocal of the product of the probabilities of selection: PSU, segment, household

x_{ihkca} = measure of the x-health characteristic of the hth person in the kth segment of the ith PSU belonging to the cth region-residence-color class and ath age-sex-color class

$\frac{n_{ik}}{n'_{ik}}$ = nonresponse adjustment

n_{ik} = the number of sample households in the kth segment of the ith PSU

n'_{ik} = the number of interviewed households in the kth segment of the ith PSU

$w'_{ih} = w_{ih} \frac{n_{ik}}{n'_{ik}}$

2. Nonresponse first-stage ratio adjusted estimate

$$x' = \sum_a \sum_c \frac{\sum_i \sum_h \sum_k w'_{ih}\, x_{ihkca}}{\sum_i P_i\, Z_{ci}} Z_c$$

$$= \sum_a \sum_c \frac{x'_{ac}}{Z'_c} Z_c$$

where

x' = the nonresponse first-stage ratio adjusted estimate of the x-health characteristic

$x'_{ac} = \sum_i \sum_h \sum_k w'_{ih}\, x_{ihkca}$

$\frac{Z_c}{Z'_c}$ = first-stage ratio adjustment

Z_c = the 1960 census population in the cth region-residence-color class

$Z'_c = \sum_i P_i\, Z_{ci}$

Z_{ci} = 1960 census figure for the cth region-residence-color class of the ith PSU

P_i = reciprocal of the probability of selecting ith PSU

3. Final post-stratified estimate

$$x'' = \sum_a \frac{\sum_c \frac{x'_{ac}}{Z'_c} Z_c}{\sum_c \frac{y'_{ac}}{Z'_c} Z_c} y_a$$

where

x'' = nonresponse two-stage ratio adjusted estimate of the x-health characteristic

$$\frac{y_a}{\sum \frac{y'_{ac}}{Z'_c} Z_c} = \text{post-stratified adjustment}$$

y_a = independent control of population count in the ath age-sex-color class

$$y'_{ac} = \sum_i \sum_h \sum_k w'_{ih} \; y_{ihkca}$$

= nonresponse adjusted estimate of the population in the acth class

y_{ihkca} = 1 if the hth person in the kth segment of the ith PSU falls in the acth class; 0 otherwise

Types of Statistics

Data collected in the Health Interview Survey are punched on cards after initial editing by hand for completeness and proper identification. There are four general types of cards—household, person, condition, and hospital. Information concerning the household such as type of living quarters, whether or not there was a phone, and data about the interview itself is punched on the household card. The person card contains the demographic characteristics for each person in the household. A condition card is prepared for each condition reported for each person in the household; information on the condition card includes whether the condition was acute or chronic, when the onset was, and any consequent limitation. A hospital card is punched for each hospital episode experienced by each person in the household; the hospital card contains dates of the hospital episodes, type of hospital, length of stay, and whether or not there was an operation.

Punched cards are transcribed to a magnetic tape and the tape is edited by a series of computer programs. During this processing, the nonresponse, first-stage, and post-stratified factors are applied to the basic weight with the newly computed weight replacing the original one. Information from each of the four types of cards is transferred to four separate tapes.

Statistics based on these units are usually divided into two types. The type of statistic depends on the period of time the respondent has to recall when he is reporting his morbidity.

Type A.—The recall period of time for prevalence and incidence data is 12 months.

Type B.—The length of the reference period for certain incidence data is 2 weeks.

Some statistics such as hospital episodes are exceptions and do not belong in either of the above classes, so they are classified as Type C. Special studies have shown that after 6 months a person's ability to report accurately on hospital episodes decreases rapidly. Therefore a 6-month reference period is used in tabulation even though the data are collected for a 12-month reference period.

The statistics are classified not only by type but also by what is called range class. The criterion for the range class is the value of the

measure of a health characteristic for an individual. These range classes stratify the statistics for variance estimation purposes. The three range classes are

Narrow range.—The measure for an individual is usually 0 or 1 and occasionally 2. This class also includes the statistics which estimate a population attribute.

Medium range.—The values for an individual are in the range 0 to 5.

Wide range.—This encompasses all statistics where a measure for an individual is greater than 5.

Statistics also vary by the form in which they are presented. They may be calculated as an aggregate, a rate, or a percentage.

Aggregates.—These are estimates of total number of events or number of persons with a given characteristic.

Rates.—The numbers of events are expressed per 100 (or any other number) persons.

Percentages.—These are estimates of the proportions with a certain attribute times 100.

Finally, the statistics can be categorized by the time interval of data collection. Quarterly, annual, and biennial estimates are published. Thus a given statistic will be of two possible types, three possible ranges, three possible forms, and refer to three possible time intervals.

The following examples illustrate some of the possible types of statistics:

Aggregates, Type B, narrow range.—The total number of acute conditions for a year in the United States or the number of injuries for all males for a year.

Rates, Type B, narrow range.—The examples in number 1 become rates if they are expressed as the number of acute conditions per person for the year and as the number of injuries per 100 males for a year.

Aggregate, Type A, medium range.—The number of chronic conditions for all people residing in the Northeast for a year.

Percentage, Type A, narrow range.—The percentage of people in the United States for a year who have one chronic condition or more.

HEALTH INTERVIEW SURVEY VARIANCES

In using estimates from sample surveys, one of the first questions asked is what is the precision of the estimator. In general there are two sources of error: (1) measurement error, which is the error that arises from response error, interviewer error, coding error, and the like, i.e., a person may not report his age accurately or an error may be made in transcribing his age to computer tape, and (2) sampling error, which is the error that is due to sampling elements from a population instead of taking a complete census. This report does not give explicit attention to measurement error, although the observant reader will note that the estimating processes described later, which yield what is termed "sampling error," do encompass a portion of measurement variance in the data but not all.

An estimate of the reliability of the results from a sample can be made from the sample itself. The form of the variance estimator depends on the sample design and the inflation procedure used. A desirable feature of the variance estimator is that it be unbiased. Another desirable feature of the estimator is that it be relatively simple to calculate.

The initial variances for the Health Interview Survey (HIS) were computed by the technique called random group method.[4] This method was used for data from runs of the 60 items collected in the calendar year 1959. The items included total males, total females, medical visits during the past 2 weeks, hospital discharges during the past 12-months, incidence of accidents, incidence of nonchronic conditions for various subclasses, and days lost from school.

To illustrate the random group procedure, consider the simplest sample design which is a

simple random sample of n observations drawn with replacement. The observations are randomly distributed among t groups consisting of $(\frac{n}{t})$ observations each. Then the estimate is

$$\hat{Var}_t\ (\bar{x}) = \sum_{i=1}^{t} (\bar{x}_i - \bar{x})^2 / t(t-1)$$

where

$\bar{x}_i$ = estimated mean of the x-health characteristic for the ith group

$\bar{x}$ = the overall estimated mean

$$= \sum_{i=1}^{t} \bar{x}_i / t$$

This random group procedure can be applied to the estimates made in HIS.

From the variance estimates for the 60 aggregate values representing each range (narrow, medium, and wide), type (2-week and 1-year reference periods), and interview period (1 quarter, 1 year, and 2 years), families of variance curves were fitted. The user of HIS data can easily obtain an average estimate of the statistic's reliability by reading the appropriate chart in Appendix I of HIS reports (PHS Pub. No. 1000—Series 10).

When the decision was made to update the charts of relative standard errors, several possibilities of estimating precision were considered. Among the methods considered were random group method, Keyfitz method, and half-sample replication method.[5] The Keyfitz method was selected for current calculations. Because of the large number of strata in the HIS design, the pseudo-replication method would require considerably more running time for calculations than the Keyfitz method. From the Keyfitz method, estimates of components of variance can be computed, whereas they cannot be obtained from the random group method and the pseudo-replication method. (This report, however, is not concerned with components of variance.) Another desirable feature of the Keyfitz method is the relative simplicity of the computations. The scheme can be used to compute estimates of variances for all stages of estimation without computing the many covariance terms separately. The basic theory, outlined in the Keyfitz article, will be stated and will be followed by the application to HIS estimator.

Basic Theory

Theorems required for derivations in subsequent sections of this report are simply stated here. The proofs of these theorems are presented in Appendix I.

Theorem 1. Variance of Estimates of Population Totals:

If x_1 and x_2 are estimates of x made from two random samples independently drawn with replacement, $\text{Var}\ (x_1 + x_2) = E\ (x_1 - x_2)^2$

Corollary 1.1:

If x_{s1} and x_{s2} are estimates of the sth stratum total x_s made from two random independent samples drawn with replacement from the sth stratum, then

$$\text{Var} \sum_s (x_{s1} + x_{s2}) = E \sum_s (x_{s1} - x_{s2})^2$$

Theorem 2. Covariance Between Two Estimates:

If x_1 and y_1 are estimates from a random sample and x_2 and y_2 are estimates from a different random sample with the two samples being drawn independently with replacement and $\text{Cov}\ (x_1, y_2) = \text{Cov}\ (x_2, y_1) = 0$ and $E\ (x_1) = E\ (x_2)$ and $E\ (y_1) = E\ (y_2)$, $\text{Cov}\ (x_1 + x_2)\ (y_1 + y_2) = E\ (x_1 - x_2)\ (y_1 - y_2)$

Corollary 2.1:

For all x_{s1} and x_{s2} mutually independent estimates of sth stratum total x_s, for all y_{s1} and y_{s2} mutually independent estimates of sth stratum total y_s, $\text{Cov}(x_{s1}, y_{s2}) = \text{Cov}(x_{s2}, y_{s1}) = 0$, and $E(x_{s1}) = E(x_{s2})$ and $E(y_{s1}) = E(y_{s2})$, then $\text{Cov} \sum_s (x_{s1} + x_{s2}), \sum_s (y_{s1} + y_{s2}) = E \sum_s (x_{s1} - x_{s2})(y_{s1} - y_{s2})$

Theorem 3. Relative Variance of Ratio Estimates:

If x_1 and y_1 are estimates from a random sample and x_2 and y_2 are estimates from a different random sample with the two samples being drawn independently at random and $\text{Cov}(x_1, y_2) = \text{Cov}(x_2, y_1) = 0$, then the relative variance (V^2) of the ratio $\left(\frac{x_1 + x_2}{y_1 + y_2}\right)$ is

$$V^2\left(\frac{x_1 + x_2}{y_1 + y_2}\right) = E\left[\frac{x_1 - x_2}{E(x_1 + x_2)} - \frac{y_1 - y_2}{E(y_1 + y_2)}\right]^2$$

Corollary 3.1:

For all x_{s1} and x_{s2} mutually independent estimates of sth stratum total x_s, for all y_{s1} and y_{s2} mutually independent estimates of sth stratum total y_s and $\text{Cov}(x_{s1}, y_{s2}) = \text{Cov}(x_{s2}, y_{s1}) = 0$, then the relative variance of the ratio $\left(\frac{\sum_s (x_{s1} + x_{s2})}{\sum_s (y_{s1} + y_{s2})}\right)$ is $V^2\left(\frac{\sum_s (x_{s1} + x_{s2})}{\sum_s (y_{s1} + y_{s2})}\right)$

$$= E \sum_s \left[\frac{x_{s1} - x_{s2}}{E \sum_s (x_{s1} + x_{s2})} - \frac{y_{s1} - y_{s2}}{E \sum_s (y_{s1} + y_{s2})}\right]^2$$

Theorem 4. Variance of Post-Stratified Estimates:

If x_{sai} are mutually independent estimates as s and i vary (x_{sai} is not independent of x_{sbi}), y_{sai} are mutually independent estimates as s and i vary (y_{sai} is not independent of y_{sbi}), $E(x_{sa1}) = E(x_{sa2})$, $E(y_{sa1}) = E(y_{sa2})$, and P_a is the precalculated total population in ath class, then the variance of the estimate

$$x = \sum_a \frac{\sum_s (x_{sa1} + x_{sa2})}{\sum_s (y_{sa1} + y_{sa2})} P_a \quad \text{is}$$

$$\text{var}(x) = E \sum_s \left[\sum_a P_a \left\{\frac{E \sum_s (x_{sa1} + x_{sa2})}{E \sum_s (y_{sa1} + y_{sa2})}\right\} \left\{\frac{x_{sa1} - x_{sa2}}{E \sum_s (x_{sa1} + x_{sa2})} - \frac{y_{sa1} - y_{sa2}}{E \sum_s (y_{sa1} + y_{sa2})}\right\}\right]^2$$

Application to HIS

Using the basic theory given in the previous section, the Keyfitz estimate of variance can be applied to the estimator used in the Health Interview Survey. The estimating equation in HIS is

$$x'' = \sum_a \frac{\sum_c \frac{x'_{ac}}{Z'_c} Z_c}{\sum_c \frac{y'_{ac}}{Z'_c} Z_c} y_a$$

In the development of the theory two independent estimates x_{s1} and x_{s2} were drawn from each stratum. In this discussion assume two PSU's are drawn independently from each stratum in the HIS sample.

Rewriting the estimating equation as

$$(1) \quad x'' = \sum_a \frac{\sum_c \frac{\sum_s (x'_{acs1} + x'_{acs2})}{\sum_s (Z'_{cs1} + Z'_{cs2})} Z_c}{\sum_c \frac{\sum_s (y'_{acs1} + y'_{acs2})}{\sum_s (Z'_{cs1} + Z'_{cs2})} Z_c} y_a$$

$$= \sum_a \frac{x'_a}{y'_a} y_a \quad \text{(post-stratified)}$$

where

x'' = published estimate of the x-health characteristic

x'_{acsi} = the nonresponse adjusted estimate of the ath age-sex-color class and cth region-residence-color class of the ith (i = 1 or 2) PSU in the sth stratum

Z'_{csi} = the 1960 census figure of the cth class of the ith PSU in sth stratum inflated by the reciprocal of the probability of selecting the ith PSU

Z_c = 1960 census population of the cth class

y'_{acsi} = the nonresponse adjusted estimate of the population of the ath age-sex-color class and cth region-residence-color class of the ith PSU in the sth stratum

y_a = independent control figure of the ath class

$$x'_a = \sum_c \frac{\sum_s (x'_{acs1} + x'_{acs2})}{\sum_s (Z'_{cs1} + Z'_{cs2})} Z_c$$

$$y'_a = \sum_c \frac{\sum_s (y'_{acs1} + y'_{acs2})}{\sum_s (Z'_{cs1} + Z'_{cs2})} Z_c$$

The variance of x'' is

$$(2)\quad \operatorname{var}(x'') = \sum_a \operatorname{var}\left(\frac{x'_a}{y'_a} y_a\right) + \sum_{a \neq b}\sum \operatorname{cov}\left(\frac{x'_a}{y'_a} y_a, \frac{x'_b}{y'_b} y_b\right)$$

The derivation of the variance of x'' is given in detail in Appendix II. By using the theorems and corollaries in Appendix I, the variance becomes

$$\operatorname{var}(x'') = \sum_a \frac{y_a^2}{(Ey'_a)^2} E \sum_s \left[\sum_c Z_c \left\{\left(\frac{E\sum_s x'_{acs}}{E\sum_s Z'_{cs}}\right)\left(\frac{\Delta x'_{acs}}{E\sum_s x'_{acs}} - \frac{\Delta Z'_{cs}}{E\sum_s Z'_{cs}}\right) - \left(\frac{Ex'_a}{Ey'_a}\right)\left(\frac{E\sum_s y'_{acs}}{E\sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{acs}}{E\sum_s y'_{acs}} - \frac{\Delta Z'_{cs}}{E\sum_s Z'_{cs}}\right)\right\}\right]^2 + \sum_{a \neq b}\sum \frac{y_a\, y_b}{Ey'_a\, Ey'_b}$$

$$E \sum_s \left[\sum_c Z_c^2 \left\{\left[\left(\frac{E\sum_s x'_{acs}}{E\sum_s Z'_{cs}}\right)\left(\frac{\Delta x'_{acs}}{E\sum_s x'_{acs}} - \frac{\Delta Z'_{cs}}{E\sum_s Z'_{cs}}\right) - \left(\frac{Ex'_a}{Ey'_a}\right)\left(\frac{E\sum_s y'_{acs}}{E\sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{acs}}{E\sum_s y'_{acs}} - \frac{\Delta Z'_{cs}}{E\sum_s Z'_{cs}}\right)\right]\left[\left(\frac{E\sum_s x'_{bcs}}{E\sum_s Z'_{cs}}\right)\left(\frac{\Delta x'_{bcs}}{E\sum_s x'_{bcs}} - \frac{\Delta Z'_{cs}}{E\sum_s Z'_{cs}}\right) - \left(\frac{Ex'_b}{Ey'_b}\right)\left(\frac{E\sum_s y'_{bcs}}{E\sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{bcs}}{E\sum_s y'_{bcs}} - \frac{\Delta Z'_{cs}}{E\sum_s Z'_{cs}}\right)\right]\right\}\right]$$

where

$$\sum_s x'_{acs} = \sum_s (x'_{acs1} + x'_{acs2})$$

$$\Delta x'_{acs} = x'_{acs1} - x'_{acs2}$$

$$(3)\quad \operatorname{var}(x'') = E\sum_s \left[\sum_a \frac{y_a}{Ey'_a} \sum_c Z_c \left\{\left(\frac{E\sum_s x'_{acs}}{E\sum_s Z'_{cs}}\right)\left(\frac{\Delta x'_{acs}}{E\sum_s x'_{acs}} - \frac{\Delta Z'_{cs}}{E\sum_s Z'_{cs}}\right) - \left(\frac{Ex'_a}{Ey'_a}\right)\left(\frac{E\sum_s y'_{acs}}{E\sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{acs}}{E\sum_s y'_{acs}} - \frac{\Delta Z'_{cs}}{E\sum_s Z'_{cs}}\right)\right\}\right]^2$$

The formula for the variance of x'' simplifies when the expected value of the y'_a's is the same as the precalculated total y_a and when the expected value of the Z'_c's is the same as Z_c. Let $x_{ac} = E \sum_s (x'_{acs1} + x'_{acs2})$ and $y_{ac} = E \sum_s (y'_{acs1} + y'_{acs2})$. The formula after these changes is

$$\text{var}(x'') = E \sum_s \left[\sum_a \sum_c \left\{ x_{ac} \left(\frac{\Delta x'_{acs}}{x_{ac}} - \frac{\Delta Z'_{cs}}{Z_c} \right) - \left(\frac{E x'_a}{y_a} \right) \left(y_{ac} \right) \left(\frac{\Delta y'_{acs}}{y_{ac}} - \frac{\Delta Z'_{cs}}{Z_c} \right) \right\} \right]^2$$

This is the true variance.

However, to obtain an estimate of variance where there is no way of calculating the true expectation, the expectations are usually dropped and the best estimates of $E\, x'_a$, x_{ac}, and y_{ac} are inserted into the equation. In HIS the best estimates of these values are the final two-stage ratio-adjusted estimates. Thus

$$(4)\quad \hat{\text{var}}(x'') = \sum_s \left[\sum_a \sum_c \left\{ x''_{ac} \left(\frac{\Delta x'_{acs}}{x''_{ac}} - \frac{\Delta Z'_{cs}}{Z_c} \right) - \left(\frac{x''_a}{y_a} \right) \left(y''_{ac} \right) \left(\frac{\Delta y'_{acs}}{y''_{ac}} - \frac{\Delta Z'_{cs}}{Z_c} \right) \right\} \right]^2$$

$$= \sum_s \left\{ \sum_a \sum_c \Delta x'_{acs} - \sum_a \sum_c x''_{ac} \frac{\Delta Z'_{cs}}{Z_c} - \sum_a \sum_c \frac{x''_a}{y_a} \Delta y'_{acs} + \sum_a \sum_c \left(\frac{x''_a}{y_a} \right) y''_{ac} \frac{\Delta Z'_{cs}}{Z_c} \right\}^2$$

$$= \sum_s \left\{ \sum_a \sum_c (x'_{acs1} - x'_{acs2}) - \sum_c \frac{Z'_{cs1} - Z'_{cs2}}{Z_c} \sum_a x''_{ac} - \sum_a \frac{x''_a}{y_a} \sum_c (y'_{acs1} - y'_{acs2}) + \sum_c \frac{Z'_{cs1} - Z'_{cs2}}{Z_c} \sum_a \frac{x''_a}{y_a} y''_{ac} \right\}^2$$

$$= \sum_s \left\{ (x'_{s1} - x'_{s2}) - \sum_c \frac{Z'_{cs1} - Z'_{cs2}}{Z_c} x''_c - \sum_a \frac{x''_a}{y_a} (y'_{as1} - y'_{as2}) + \sum_c \frac{Z'_{cs1} - Z'_{cs2}}{Z_c} \sum_a \frac{x''_a}{y_a} y''_{ac} \right\}^2$$

$$= \sum_s \left\{ (x'_{s1} - x'_{s2}) - \sum_a \frac{x''_a}{y_a} (y'_{as1} - y'_{as2}) - \sum_c \frac{Z'_{cs1} - Z'_{cs2}}{Z_c} (x''_c - \sum_a \frac{x''_a}{y_a} y''_{ac}) \right\}^2$$

where

x'' = the published HIS estimate of the x-health characteristic

x'_{si} = the simple nonresponse adjusted estimate of the health-characteristic for ith (i=1 or 2) PSU in the sth stratum

x''_a = HIS final estimate for the ath age-sex-color class

y_a = independent control figure of the ath class

y'_{asi} = nonresponse adjusted estimate of the population for the ath age-sex-color class in the sith PSU

Z'_{csi} = nonresponse adjusted estimate of the population of the cth color-residence-region class in sith PSU

Z_c = 1960 census population of the cth class

x''_c = HIS final estimate for cth color-residence-region class

y''_{ac} = final estimate of the population in the acth cell

In the preceding paragraphs the formula for calculating sampling errors for the estimator used in HIS has been developed. In implementing this formula, however, difficulties arose since all the basic assumptions were not met by the sample design of HIS. (Programming problems are discussed in Appendix III.)

First of all, in the Health Interview Survey only one PSU is drawn from each of 357 strata. Some strata are PSU's and thus have a first stage probability of selection equal to one (called self-representing PSU's—SR PSU's). In the remaining strata the PSU's have a first stage probability of selection not equal to one (called non-self-representing PSU's—NSR PSU's). This difference in the first stage of probability of selection causes the estimation procedure to be slightly dissimilar for the two types of PSU's. This in turn causes the variance formulas to differ.

The final estimate in HIS can be expressed as:

$$x'' = x''_{NSR} + x''_{SR}$$

and therefore

$$\hat{Var}(x'') = \hat{Var}(x''_{SR}) + \hat{Var}(x''_{NSR})$$

Self-Representing PSU's

The estimating equation given for HIS was

$$x'' = \sum_a \frac{\sum_c \frac{x'_{ac}}{Z'_c} Z_c}{\sum_c \frac{y'_{ac}}{Z'_c}} y_a$$

where

$Z'_c = \sum_i P_i Z_{ci}$

Z_{ci} = 1960 census figure for the cth region-residence-color class of the ith PSU

P_i = reciprocal of the probability of selecting ith PSU

For SR PSU's, $P_i = 1$; therefore, $Z'_c = Z_c$. The equation then becomes

$$x'' = \sum_a \frac{\sum_c x'_{ac}}{\sum_c y'_{ac}} y_a$$

$$= \sum_a \frac{x'_a}{y'_a} y_a$$

or rewriting

$$x''_{SR} = \sum_a \frac{\sum_s (x'_{as1} + x'_{as2})}{\sum_s (y'_{as1} + y'_{as2})} y_a$$

where

x''_{SR} = the HIS regular estimate for SR PSU's

x'_{asi} = the nonresponse adjusted estimate of the ath age-sex-color class of the i (i = 1 or 2) segment in the sth stratum

y'_{asi} = the nonresponse adjusted estimate of the population of the ath age-sex-color class of the ith segment in the sth stratum

y_a = independent control figure of the ath class

Then from the previous formula (4) developed

$$(5)\ \hat{var}(x''_{SR}) = \sum_s \left\{ (x'_{s1} - x'_{s2}) - \sum_a \frac{x''_a}{y_a} (y'_{as1} - y'_{as2}) \right\}^2$$

The first problem encountered in implementing the formula for the variance estimator is that each strata represents just one PSU. One solution might be to collapse strata and pair PSU's, but this would introduce a component of the variance arising from sampling PSU's. This is unrealistic since the PSU's come into the sample with certainty and estimates only have a within PSU component of variance. Pairing of segments within PSU's, however, would produce the necessary two observations without adding between-PSU variation.

Before selecting the units into the sample, the segments are ordered by geographical location and then systematically sampled. This systematic sample stratifies the segments into strata with one segment chosen from each stratum. In order to apply the Keyfitz procedure, adjacent strata were collapsed, resulting in a pair of segments. In this pairing of segments, it is assumed $Ex_1 = Ex_2$. Although the Keyfitz procedure is essentially unbiased, it is biased when the technique of collapsing strata is used. It is believed that this bias is negligible relative to the variance.

Non-Self-Representing PSU's

For these PSU's the estimating equation is

$$x''_{NSR} = \sum_a \frac{\sum_c \frac{\sum_s (x'_{acs1} + x'_{acs2})}{\sum_s (Z'_{cs1} + Z'_{cs2})} Z_c}{\sum_c \frac{\sum_s (y'_{acs1} + y'_{acs2})}{\sum_s (Z'_{cs1} + Z'_{cs2})} Z_c} y_a$$

Here all symbols are defined as they were before. The variance formula is

$$(6)\ \hat{\text{var}}\,(x''_{NSR}) = \sum_s \left\{ (x'_{s1} - x'_{s2}) - \sum_a \frac{x''_a}{y_a}(y'_{as1} - y'_{as2}) - \sum_c \frac{Z'_{cs1} - Z'_{cs2}}{Z_c}\left(x''_c - \sum_a \frac{x''_a}{y_a} y''_{ac}\right)\right\}^2$$

In this situation there is only one PSU from each stratum in the sample. However, the strata consisted of more than one PSU so that there is between-PSU variance. Thus the collapsed strata technique described by Hansen, Hurwitz, and Madow[6] is applicable. The procedure is to pair the strata and then treat the two PSU's in the sample as two observations from the same stratum. The important thing is to pair the strata so the gains through this restratification are at a minimum. This should be done independently of the sample results. Unfortunately it cannot be assumed that the expected sizes are the same, so an adjustment is made as follows:

$$P_{si} = \frac{\text{1960 population of the sith NSR stratum}}{\text{1960 population of the sth pair of NSR stratum}}$$

The modification becomes[7]

$$(7)\ \hat{\text{var}}\,(x''_{NSR}) = \sum_s \left\{ (2P_{s2}\,x'_{s1} - 2P_{s1}\,x'_{s2}) - \sum_a \frac{x''_a}{y_a}(2P_{s2}\,y'_{as1} - 2P_{s1}\,y'_{as2}) - \sum_c \frac{2P_{s2}\,Z'_{cs1} - 2P_{s1}\,Z'_{cs2}}{Z_c}\left(x''_c - \sum_a \frac{x''_a}{y_a} y''_{ac}\right)\right\}^2$$

ESTIMATED VARIANCES

The time and cost of computing estimates of variances for each statistic published in HIS by the Keyfitz method would be prohibitive. Also the report would be greatly complicated if variances for all means and aggregates for the total sample and variances for all classes were presented. Instead of presenting variances for each statistic, the data can be grouped and "average variances" given. The term "average variance" here means that estimates of variances are computed for selected statistics of a group

and these variance estimates are used to provide the reliability for any statistics belonging to that group. Grouping the statistics is not easy, but two points to consider are that the survey characteristics such as prevalence of any diseases represented in a group should have similar design effects and that the groups should cover the possible range of variation of the data. The grouping of HIS data is done by considering each type (Type A and Type B) per range (narrow, medium, and wide) per time interval (quarter, year, and 2 years) as a group.

Empirically, it has been shown that there is a relationship between the size of the estimate and the estimate's relative variance; as the estimate increases in size its relative variance decreases. The relationship is expressed by the formula

$$V_x^2 = a + \frac{b}{x''}$$

Therefore using the relative variances of the selected statistics of a group, values for a and b are calculated. Then a smooth relative standard error curve in percent can be drawn. From this curve the reliability of any estimate falling into the group can be determined.

The standard method of estimating a and b is the method of least squares. The least squares estimators give values of a and b which minimize the sum of squares of deviations between the observed values $V_x^{2\prime\prime}$ and the predictions V_x^2. Thus

$$S = \sum_i \left[V_{x_i}^{2\prime\prime} - a - b/x_i''\right]^2$$

is minimized. The exact estimators which will minimize S are found by differentiating the sum with respect to a and b and equating to zero. However, the method used here was to minimize the squared relative residuals of V_x^2. Here the quantity

$$S' = \sum_i \left[\frac{V_{x_i}^{2\prime\prime} - a - b/x_i''}{V_{x_i}^2}\right]^2$$

is the one to be minimized. Upon differentiating S' with respect to a and b and equating to zero, formulas for a and b are found.

These formulas have the term $V_{x_i}^2$ in them. Since this is unknown, an iterative procedure is employed. Substituting $V_{x_i}^{2\prime\prime}$ for $V_{x_i}^2$, values a_1 and b_1 are computed and then used to calculate $\hat{V}_{x_i}^2 = a_1 + b_1/x_i''$. Next, new estimates, a_2 and b_2, are figured but this time using $\hat{V}_{x_i}^2$ for $V_{x_i}^2$. If $\left|\frac{a_2 - a_1}{a_2}\right| \geqslant 2\%$ or $\left|\frac{b_2 - b_1}{b_2}\right| \geqslant 2\%$, the process is repeated with $\hat{V}_{x_i}^2 = a_2 + b_2/x_i''$ replacing $V_{x_i}^2$. Iterations are run until $\left|\frac{a_j - a_{j-1}}{a_j}\right| \leqslant 2\%$ and $\left|\frac{b_j - b_{j-1}}{b_j}\right| \leqslant 2\%$.

Estimates of incidence of acute conditions for classes are given in the table with the estimated relative variances as calculated by the Keyfitz method and the smoothed relative variances as computed from the fitted curves. These estimates constitute one group from HIS data. They are representatives of Type B, narrow range, annual statistics. Column (3) was calculated by the Keyfitz method explained in the preceding sections. Columns (2) and (3) were input data into a computer program that computed estimates of a and b by the process discussed above. The output from the program is in column (4), and value a = 0.000280 and b = 42,522.883325. Using a and b, the curve given in figure 1 was drawn.

The table shows that as the estimates increase in size their relative variances decrease. For example, the smallest estimate, 1,064,000, has a relative variance of .044017, while the largest estimate, 387,358,000, has a relative variance of only .000343. For the smaller estimates the smoothed relative variance is actually greater than the computed value. Not until the size of the estimate is around 60,000

Incidence of acute conditions for classes with their actual and smoothed relative variances: United States, July 1963-June 1964

Class (1)	Estimate (in thousands) (2)	Relative variance	
		Actual (3)	Smooth (4)
Number of			
Other respiratory conditions—male, 15-44	1,064	.044017	.057333
Digestive system conditions—male, 45 and over	2,012	.019628	.030473
Digestive system conditions—female, 15-44	4,827	.009058	.012886
Injuries—female, 45 and over	5,702	.007866	.010957
All acute conditions—male, retired, 45 and over	6,269	.005896	.009996
Cases of influenza—female, 45 and over	6,408	.006188	.009786
Injuries—female, 15-44	8,402	.005229	.007539
Injuries—male, 25-44	8,441	.006982	.007505
Other respiratory conditions—both sexes, all ages	8,524	.005627	.007435
Digestive system conditions—male, all ages	9,961	.003390	.006408
Contusions and superficial injuries—both sexes, all ages	10,421	.004303	.006139
Other current injuries—both sexes, all ages	12,603	.002849	.005131
Open wounds and lacerations—both sexes, all ages	15,835	.002650	.004148
Fractures, dislocations, sprains, and strains—both sexes, all ages	16,366	.002838	.004024
Digestive system conditions—both sexes, all ages	20,608	.002135	.003261
Upper respiratory conditions—female, 17-44	21,572	.001949	.003129
All other acute conditions—both sexes, all ages	51,941	.001108	.001485
All acute conditions—45-64	52,539	.001071	.001471
Infective and parasitic diseases—both sexes, all ages	55,283	.001491	.001414
All acute conditions—15-24	55,836	.001053	.001403
Cases of influenza—both sexes, all ages	61,980	.001948	.001296
All acute conditions—under 5	76,083	.000819	.001114
All acute conditions—45-64	79,701	.000747	.001078
All acute conditions—5-14	103,653	.000686	.000902
All acute conditions—currently employed, 17 and over	104,100	.000619	.000899
Upper respiratory conditions—both sexes, all ages	133,797	.000554	.000770
All acute conditions—male	180,182	.000455	.000653
All acute conditions—female	207,175	.000428	.000609
All acute conditions—both sexes, all ages	387,358	.000343	.000473

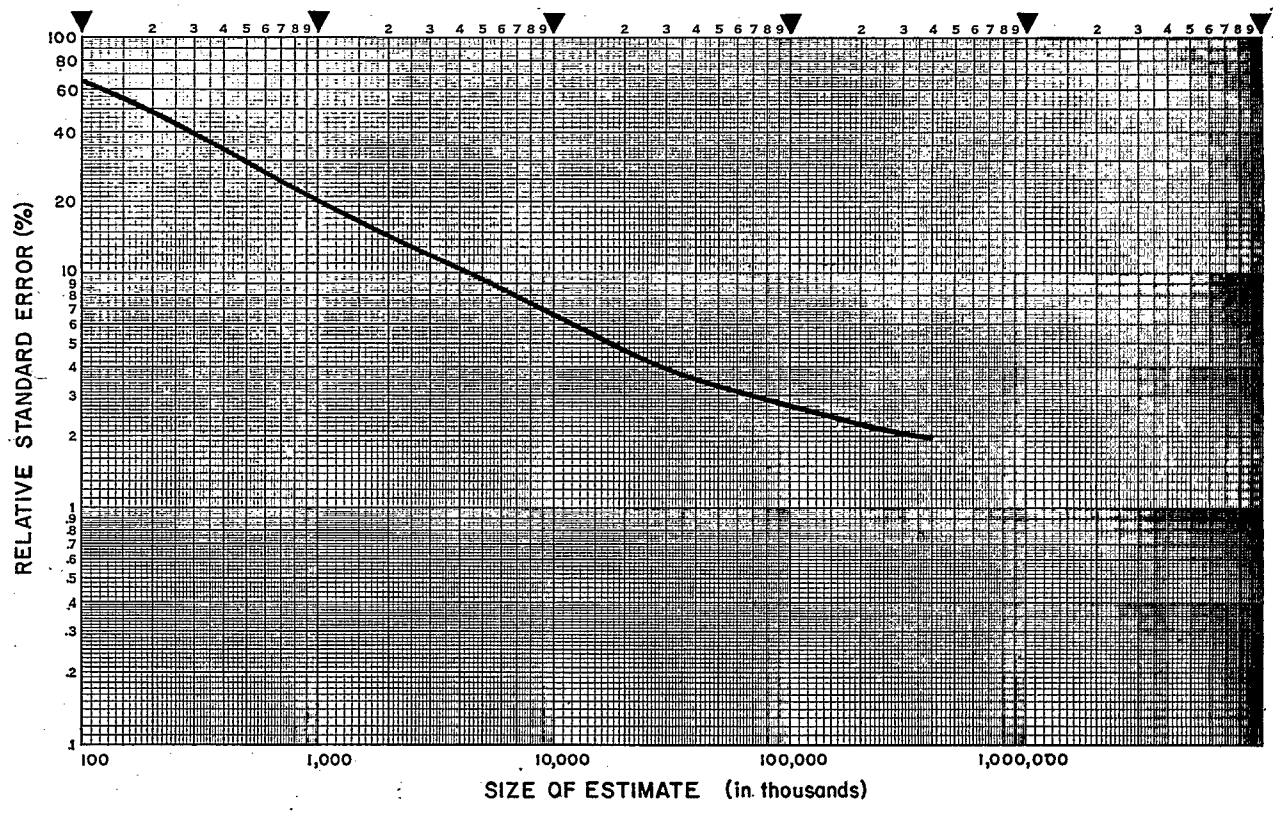

Figure 1. Smoothed relative standard error curve for aggregate estimates based on 1 year of data collection for Type B data, narrow range.

does the smoothed relative variance drop below the calculated value.

Figures 2 and 3 are more examples of the smoothed relative standard error curves that result from Keyfitz computed relative variances and the curve fitting computer program. To illustrate the use of these curves, let's say you have an estimate of 20,000,000. If the estimate is number of acute conditions, you would read a relative standard error of 4.9 percent from figure 1. For number of dental visits, using figure 2, the relative standard error is 6.1 percent. The error is 2 percent from figure 3 if the estimate is number of chronic conditions. To convert from relative standard error to standard error, multiply the estimate times the relative standard error in percent. For the above figures the standard errors are 980,000 (4.9 percent of 20 million), 1,220,000 (6.1 percent of 20 million), and 400,000 (2 percent of 20 million).

Curves necessary to find the reliability of any of the estimates given in an HIS report (PHS Pub. No. 1000—Series 10) may be found in Appendix I of the report.

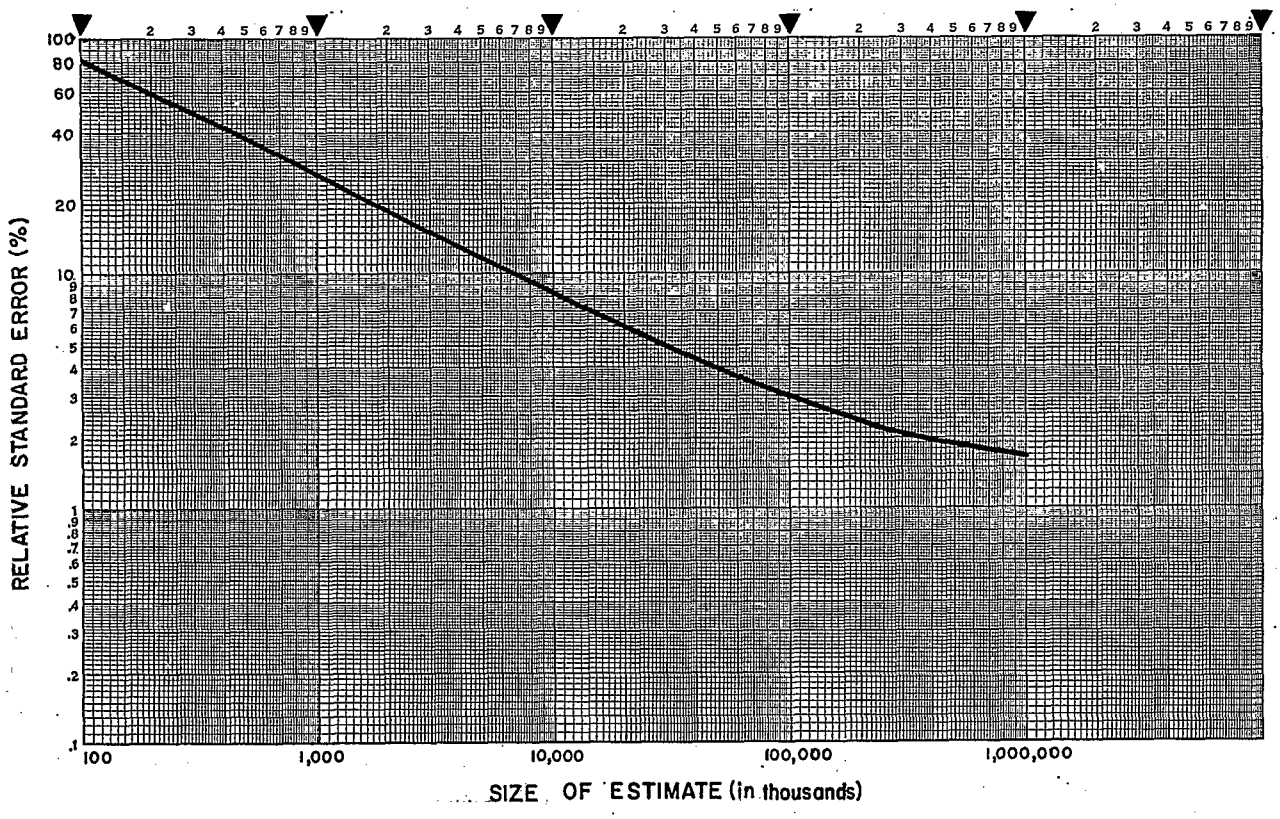

Figure 2. Smoothed relative standard error curve for aggregate estimates based on 1 year of data collection for Type B data, medium range.

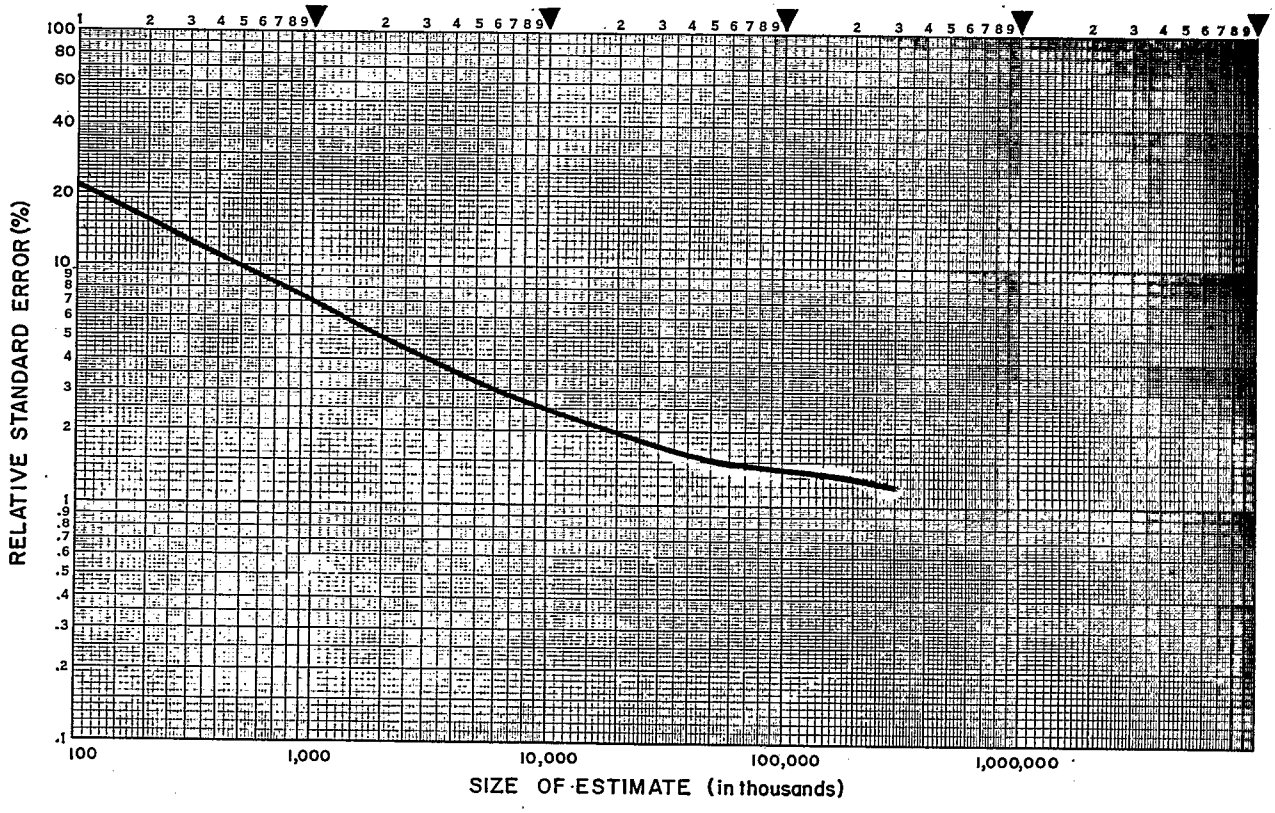

Figure 3. Smoothed relative standard error curve for aggregate estimates based on 1 year of data collection for Type A data, medium range.

REFERENCES

[1] Keyfitz, N.: Estimates of sampling variance where two units are selected from each stratum. *J.Am. statist.Ass.* 52(280):503-510, Dec. 1957.

[2] U.S. Bureau of the Census: *The Current Population Survey—a Report on Methodology.* Technical Paper No. 7. Washington. U.S. Government Printing Office, 1963.

[3] National Center for Health Statistics: Health Survey procedure. *Vital and Health Statistics.* PHS Pub. No. 1000-Series 1-No. 2. Public Health Service. Washington. U.S. Government Printing Office, May 1964.

[4] National Health Survey: The statistical design of the Health Household-Interview Survey. *Health Statistics.* PHS Pub. No. 584-A2. Public Health Service. Washington. U.S. Government Printing Office, July 1958.

[5] National Center for Health Statistics: Replication, an approach to the analysis of data from complex surveys. *Vital and Health Statistics.* PHS Pub. No. 1000-Series 2-No. 14. Public Health Service. Washington. U.S. Government Printing Office, Apr. 1966.

[6] Hansen, M. H., Hurwitz, W. N., and Madow, W. G.: *Sample Survey Methods and Theory.* New York. John Wiley and Sons, Inc., 1953. Vol. I, pp. 399-401 and 419.

[7] Ibid., Vol. II, pp. 218-222.

[8] Ibid., pp. 107-109.

———— ○○○ ————

APPENDIX I

PROOFS OF BASIC THEORY

Theorem 1. Variance of Estimates of Population Totals:

If x_1 and x_2 are estimates of x made from two random samples independently drawn with replacement, Var $(x_1 + x_2) = E (x_1 - x_2)^2$

Proof:

$$E (x_1) = E (x_2) \quad \text{given}$$

$$E (x_1 \; x_2) = E (x_1) \, E (x_2) \quad \text{given}$$

$$\text{Var } (x_1 + x_2) = \text{Var } (x_1) + \text{Var } (x_2)$$

$$= E (x_1^2) - [E(x_1)]^2 + E (x_2^2) - [E(x_2)]^2$$

$$= E (x_1^2) + E (x_2^2) - 2 E (x_1 \; x_2)$$

$$= E (x_1 - x_2)^2$$

Corollary 1.1:

If x_{s1} and x_{s2} are estimates of the sth stratum total x_s made from two random independent samples drawn with replacement from the sth stratum, then

$$\text{Var } \sum_s (x_{s1} + x_{s2}) = E \sum_s (x_{s1} - x_{s2})^2$$

Proof:

$$E (x_{s1}) = E (x_{s2})$$

$$E (x_{s1} \; x_{s2}) = E (x_{s1}) \, E (x_{s2})$$

$$\text{Cov } [(x_{s1} + x_{s2}), (x_{j1} + x_{j2})] = 0$$

$$\text{Var } \sum_s (x_{s1} + x_{s2}) = \sum_s \text{Var } (x_{s1}) + \sum_s \text{Var } (x_{s2})$$

$$= \sum_s (E (x_{s1}^2) - [E (x_{s1})]^2 + E (x_{s2}^2) - [E(x_{s2})]^2)$$

$$= \sum_s (E (x_{s1}^2) + E (x_{s2}^2) - 2 E (x_{s1} \; x_{s2}))$$

$$= \sum_s E (x_{s1} - x_{s2})^2$$

$$= E \sum_s (x_{s1} - x_{s2})^2$$

Theorem 2. Covariance Between Two Estimates:

If x_1 and y_1 are estimates from a random sample and x_2 and y_2 are estimates from a different random sample with the two samples being drawn independently with replacement and Cov (x_1, y_2) = Cov $(x_2, y_1) = 0$ and $E (x_1) = E (x_2)$ and $E (y_1) = E (y_2)$, Cov $(x_1 + x_2)$ $(y_1 + y_2) = E (x_1 - x_2) (y_1 - y_2)$

Proof:

$$\text{Cov}(x_1 + x_2)(y_1 + y_2) = \text{Cov}(x_1 y_1) + \text{Cov}(x_2 y_2)$$

$$= E(x_1 y_1) - E(x_1) E(y_1) + E(x_2 y_2) - E(x_2) E(y_2)$$

$$= E(x_1 y_1) - E(x_1) E(y_2) + E(x_2 y_2) - E(x_2) E(y_1)$$

$$= E(x_1 y_1) - E(x_1 y_2) + E(x_2 y_2) - E(x_2 y_1)$$

$$= E(x_1 - x_2)(y_1 - y_2)$$

Corollary 2.1:

For all x_{s1} and x_{s2} mutually independent estimates of sth stratum total x_s, for all y_{s1} and y_{s2} mutually independent estimates of sth stratum total y_s, $\text{Cov}(x_{s1}, x_{s2}) = \text{Cov}(y_{s1}, y_{s2}) = 0$, and $E(x_{s1}) = E(x_{s2})$ and $E(y_{s1}) = E(y_{s2})$, then $\text{Cov}\left[\sum_s (x_{s1} + x_{s2}), \sum_s (y_{s1} + y_{s2})\right] = E \sum_s (x_{s1} - x_{s2})(y_{s1} - y_{s2})$

Proof:

$$\text{Cov}\left[\sum_s (x_{s1} + x_{s2}), \sum_s (y_{s1} + y_{s2})\right]$$

$$= \sum_s \text{Cov}(x_{s1} + x_{s2}, y_{s1} + y_{s2}) + \sum_{s \neq j} \sum (x_{s1} + x_{s2}, y_{j1} + y_{s2})$$

$$= \sum_s E(x_{s1} - x_{s2})(y_{s1} - y_{s2})$$

$$= E \sum_s (x_{s1} - x_{s2})(y_{s1} - y_{s2})$$

Theorem 3. Relative Variance of Ratio Estimates:

If x_1 and y_1 are estimates from a random sample and x_2 and y_2 are estimates from a different random sample with the two samples being drawn independently at random and $\text{Cov}(x_1, y_2) = \text{Cov}(x_2, y_1) = 0$, then the relative variance (V^2) of the ratio

$$\left(\frac{x_1 + x_2}{y_1 + y_2}\right) \text{ is}$$

$$V^2\left(\frac{x_1 + x_2}{y_1 + y_2}\right) = E\left[\frac{x_1 - x_2}{E(x_1 + x_2)} - \frac{y_1 - y_2}{E(y_1 + y_2)}\right]^2$$

Proof:[8]

$$V^2\left(\frac{x_1 + x_2}{y_1 + y_2}\right) = V^2(x_1 + x_2) + V^2(y_1 + y_2) - 2V(x_1 + x_2)(y_1 + y_2)$$

$$= \frac{\text{Var}(x_1 + x_2)}{[E(x_1 + x_2)]^2} + \frac{\text{Var}(y_1 + y_2)}{[E(y_1 + y_2)]^2} - 2\frac{\text{Cov}(x_1 + x_2)(y_1 + y_2)}{E(x_1 + x_2) E(y_1 + y_2)} \text{ (definition)}$$

$$= \frac{E(x_1 - x_2)^2}{[E(x_1 + x_2)]^2} + \frac{E(y_1 - y_2)^2}{[E(x_1 + y_2)]^2} - 2\frac{E(x_1 - x_2)(y_1 - y_2)}{E(x_1 + x_2) E(y_1 + y_2)} \text{ (using theorem 1 and theorem 2)}$$

$$= E\left[\frac{(x_1 - x_2)^2}{[E(x_1 + x_2)]^2} - \frac{2(x_1 - x_2)(y_1 - y_2)}{E(x_1 + x_2) E(y_1 + y_2)} + \frac{(y_1 - y_2)^2}{[E(y_1 + y_2)]^2}\right]$$

$$= E\left[\frac{x_1 - x_2}{E(x_1 + x_2)} - \frac{y_1 - y_2}{E(y_1 + y_2)}\right]^2$$

Corollary 3.1:

For all x_{s1} and x_{s2} mutually independent estimates of sth stratum total x_s, for all y_{s1} and

y_{s2} mutually independent estimates of sth stratum total y_s and Cov (x_{s1}, y_{s2}) = Cov $(x_{s2}, y_{s1}) = 0$, then the relative variance of the ratio

$$\left(\frac{\sum_s (x_{s1} + x_{s2})}{\sum_s (y_{s1} + y_{s2})}\right) \text{ is}$$

$$V^2\left(\frac{\sum_s (x_{s1} + x_{s2})}{\sum_s (y_{s1} + y_{s2})}\right) = E \sum_s \left[\frac{x_{s1} - x_{s2}}{E \sum_s (x_{s1} + x_{s2})} - \frac{y_{s1} - y_{s2}}{E \sum_s (y_{s1} + y_{s2})}\right]^2$$

Proof:

This proof is the same as the proof of Theorem 3.

$$V^2\left(\frac{\sum_s (x_{s1} + x_{s2})}{\sum_s (y_{s1} + y_{s2})}\right)$$

$$\doteq V^2 \sum_s (x_{s1} + x_{s2}) + V^2 \sum_s (y_{s1} + y_{s2}) - 2V \sum_s (x_{s1} + x_{s2}) \sum_s (y_{s1} + y_{s2})$$

$$= \frac{E \sum_s (x_{s1} - x_{s2})^2}{[E \sum_s (x_{s1} + x_{s2})]^2} + \frac{E \sum_s (y_{s1} - y_{s2})^2}{[E \sum_s (y_{s1} + y_{s2})]^2} - 2 \frac{E \sum_s (x_{s1} - x_{s2})(y_{s1} - y_{s2})}{E \sum_s (x_{s1} + x_{s2}) E \sum_s (y_{s1} + y_{s2})}$$

$$= E \sum_s \left[\frac{x_{s1} - x_{s2}}{E \sum_s (x_{s1} + x_{s2})} - \frac{y_{s1} - y_{s2}}{E \sum_s (y_{s1} + y_{s2})}\right]^2$$

Theorem 4. Variance of Post-Stratified Estimates:

If x_{sai} are mutually independent estimates as s and i vary (x_{sai} is not independent of x_{sbi}), y_{sai} are mutually independent estimates as s and i vary (y_{sai} is not independent of y_{sbi}), E (x_{sa1}) = E (x_{sa2}), E (y_{sa1}) = E (y_{sa2}), and P_a is the pre-calculated total population in ath class, then the variance of the estimate

$$x = \sum_a \frac{\sum_s (x_{sa1} + x_{sa2})}{\sum_s (y_{sa1} + y_{sa2})} P_a \text{ is}$$

$$\text{Var}(x) = E \sum_s \left[\sum_a P_a \left\{\frac{E \sum_s (x_{sa1} + x_{sa2})}{E \sum_s (y_{sa1} + y_{sa2})}\right\} \left\{\frac{x_{sa1} - x_{sa2}}{E \sum_s (x_{sa1} + x_{sa2})} - \frac{y_{sa1} - y_{sa2}}{E \sum_s (y_{sa1} + y_{sa2})}\right\}\right]^2$$

Proof:

The proof will be given for the simple case of only two different a classes; therefore $x = x_a + x_b$ where

$$x_a = \frac{\sum_s (x_{as1} + x_{as2})}{\sum_s (y_{as1} + y_{as2})} P_a \text{ and } x_b = \frac{\sum_s (x_{bs1} + x_{bs2})}{\sum_s (y_{bs1} + y_{bs2})} P_b.$$

The generalization can easily be made to the case where there are k different a classes.

$$\text{Var}(x) = \text{Var}(x_a) + \text{Var}(x_b) + 2 \text{ Cov}(x_a, x_b)$$

The proof of this theorem will be in three parts: (1) variance of x_a,

(2) covariance $\left(\frac{\sum_s (x_{as1} + x_{as2})}{\sum_s (y_{as1} + y_{as2})} P_a, \frac{\sum_s (x_{bs1} + x_{bs2})}{\sum_s (y_{bs1} + y_{bs2})} P_b\right)$

and (3) covariance $\left(\frac{\sum_s (x_{as1} + x_{as2})}{\sum_s (y_{as1} + y_{as2})}, \frac{\sum_s (x_{bs1} + x_{bs2})}{\sum_s (y_{bs1} + y_{bs2})}\right)$

Part 1—Variance of x_a

$$x_a = \frac{\sum_s (x_{as1} + x_{as2})}{\sum_s (y_{as1} + y_{as2})} P_a$$

The P_a is treated as a constant with x_{asi} and y_{asi} being the random variables. The variance will be derived by using Corollary 3.1.

$$\text{Var}(x_a) = P_a^2 \text{ Var} \left(\frac{\sum_s (x_{as1} + x_{as2})}{\sum_s (y_{as1} + y_{as2})}\right)$$

Let

$$\sum_s (x_{as1} + x_{as2}) = \sum_s x_{as}$$

$$\sum_s (y_{as1} + y_{as2}) = \sum_s y_{as}$$

From Corollary 3.1

$$V^2 \left(\frac{\sum_s (x_{as1} + x_{as2})}{\sum_s (y_{as1} + y_{as2})}\right) = E \sum_s \left[\frac{x_{as1} - x_{as2}}{E \sum_s x_{as}} - \frac{y_{as1} - y_{as2}}{E \sum_s y_{as}}\right]^2$$

Thus changing relative variances into variance.

$$V^2 \left(\frac{\sum_s x_{as}}{\sum_s y_{as}}\right) = \frac{\text{Var}\left(\frac{\sum_s x_{as}}{\sum_s y_{as}}\right)}{\left(\frac{E \sum_s x_{as}}{E \sum_s y_{as}}\right)^2}$$

$$\text{Var}\left(\frac{\sum_s x_{as}}{\sum_s y_{as}}\right) = \left(\frac{E \sum_s x_{as}}{E \sum_s y_{as}}\right)^2 E \sum_s \left[\frac{x_{as1} - x_{as2}}{E \sum_s x_{as}} - \frac{y_{as1} - y_{as2}}{E \sum_s y_{as}}\right]^2$$

$$= E \sum_s \left[\frac{E \sum_s x_{as}}{E \sum_s y_{as}} \left\{\frac{x_{as1} - x_{as2}}{E \sum_s x_{as}} - \frac{y_{as1} - y_{as2}}{E \sum_s y_{as}}\right\}\right]^2$$

Now going back to

$$\text{Var}(x_a) = P_a^2 \text{ Var}\left(\frac{\sum_s x_{as}}{\sum_s y_{as}}\right)$$

$$= P_a^2 \; E \sum_s \left[\frac{E \sum_s x_{as}}{E \sum_s y_{as}} \left\{\frac{x_{as1} - x_{as2}}{E \sum_s x_{as}} - \frac{y_{as1} - y_{as2}}{E \sum_s y_{as}}\right\}\right]^2$$

Similarly, this is true for variance of x_b.

Part 2—Covariance $\left(\frac{\sum_s x_{as}}{\sum_s y_{as}} P_a, \frac{\sum_s x_{bs}}{\sum_s y_{bs}} P_b\right)$

$$\text{Cov}\left(\frac{\sum_s x_{as}}{\sum_s y_{as}} P_a, \frac{\sum_s x_{bs}}{\sum_s y_{bs}} P_b\right)$$

$$= E\left[\left(\frac{\sum_s x_{as}}{\sum_s y_{as}} P_a - E \frac{\sum_s x_{as}}{\sum_s y_{as}} P_a\right) \left(\frac{\sum_s x_{bs}}{\sum_s y_{bs}} - E \frac{\sum_s x_{bs}}{\sum_s y_{bs}} P_b\right)\right]$$

$$= P_a P_b \text{ Cov}\left(\frac{\sum_s x_{as}}{\sum_s y_{as}}, \frac{\sum_s x_{bs}}{\sum_s y_{bs}}\right)$$

Part 3—Covariance $\left(\frac{\sum_s x_{as}}{\sum_s y_{as}}, \frac{\sum_s x_{bs}}{\sum_s y_{bs}}\right)$

$$E\frac{\sum_s x_{as}}{\sum_s y_{as}} \doteq \frac{E\sum_s x_{as}}{E\sum_s y_{as}}$$

For ease of notation let

$$\frac{E\sum_s x_{as}}{E\sum_s y_{as}} = R_a \text{ and } \frac{E\sum_s x_{bs}}{E\sum_s y_{bs}} = R_b$$

$$\text{Cov}\left(\frac{\sum_s x_{as}}{\sum_s y_{as}}, \frac{\sum_s x_{bs}}{\sum_s y_{bs}}\right)$$

$$= E\left[\left(\frac{\sum_s x_{as}}{\sum_s y_{as}} - R_a\right)\left(\frac{\sum_s x_{bs}}{\sum_s y_{bs}} - R_b\right)\right]$$

$$= E\left[\left(\frac{\sum_s x_{as}}{\sum_s y_{as}} - \frac{R_a \sum_s y_{as}}{\sum_s y_{as}}\right)\left(\frac{\sum_s x_{bs}}{\sum_s y_{bs}} - \frac{R_b \sum_s y_{bs}}{\sum_s y_{bs}}\right)\right]$$

$$= E\left[\left(\frac{\sum_s x_{as} - R_a \sum_s y_{as}}{\sum_s y_{as}}\right)\left(\frac{\sum_s x_{bs} - R_b \sum_s y_{bs}}{\sum_s y_{bs}}\right)\right]$$

$$= E\left[\left(\frac{\sum_s x_{as} - R_a \sum_s y_{as} - E\sum_s x_{as} + R_a \sum_s y_{as}}{\sum_s y_{as}}\right)\right.$$

$$\left.\left(\frac{\sum_s x_{bs} - R_b \sum_s y_{bs} - E\sum_s x_{bs} + R_b \sum_s y_{bs}}{\sum_s y_{bs}}\right)\right]$$

Let $\sum_s x_{as} - E\sum_s x_{as} = \delta_{\sum_s x_{as}}$

and $\sum_s y_{as} - E\sum_s y_{as} = \delta_{\sum_s y_{as}}$

Then have

$$\text{Cov}\left(\frac{\sum_s x_{as}}{\sum_s y_{as}}, \frac{\sum_s x_{bs}}{\sum_s y_{bs}}\right) = E\left[\left(\frac{\delta_{\sum_s x_{as}} - R_a \delta_{\sum_s y_{as}}}{\sum_s y_{as}}\right)\right.$$

$$\left.\left(\frac{\delta_{\sum_s x_{bs}} - R_b \delta_{\sum_s y_{bs}}}{\sum_s y_{bs}}\right)\right]$$

$$= E\left[\left(\frac{\delta_{\sum_s x_{as}} - R_a \delta_{\sum_s y_{as}}}{\sum_s y_{as}}\right)\left(\frac{\sum_s y_{as}}{E\sum_s y_{as} + \delta_{\sum_s y_{as}}}\right)\right.$$

$$\left.\left(\frac{\delta_{\sum_s x_{bs}} - R_b \delta_{\sum_s y_{bs}}}{\sum_s y_{bs}}\right)\left(\frac{\sum_s y_{bs}}{E\sum_s y_{bs} + \delta_{\sum_s y_{bs}}}\right)\right]$$

$$= E\left[\left(\frac{\delta_{\sum_s x_{as}} - R_a \delta_{\sum_s y_{as}}}{E\sum_s y_{as}}\right)\left(\frac{\delta_{\sum_s y_{bs}} - R_b \delta_{\sum_s y_{bs}}}{E\sum_s y_{bs}}\right)\right.$$

$$\left.\left(\frac{1}{1 + \frac{\delta_{\sum_s y_{as}}}{E\sum_s y_{as}}}\right)\left(\frac{1}{1 + \frac{\delta_{\sum_s y_{bs}}}{E\sum_s y_{bs}}}\right)\right.$$

Expanding the last terms,

$$\frac{1}{1 + \frac{\delta_{\sum_s y_{as}}}{E\sum_s y_{as}}} = 1 - \left(\frac{\delta_{\sum_s y_{as}}}{E\sum_s y_{as}}\right) + \left(\frac{\delta_{\sum_s y_{as}}}{E\sum_s y_{as}}\right)^2 - \cdots$$

$$\frac{1}{1 + \frac{\delta_{\sum_s y_{bs}}}{E\sum_s y_{bs}}} = 1 - \left(\frac{\delta_{\sum_s y_{bs}}}{E\sum_s y_{bs}}\right) + \left(\frac{\delta_{\sum_s y_{bs}}}{E\sum_s y_{bs}}\right)^2 - \cdots$$

Disregarding all but the first terms,

$$\text{Cov}\left(\frac{\sum_s x_{as}}{\sum_s y_{as}}, \frac{\sum_s x_{bs}}{\sum_s y_{bs}}\right) \doteq E\left[\left(\frac{\delta_{\sum_s x_{as}} - R_a \delta_{\sum_s y_{as}}}{E\sum_s y_{as}}\right)\right.$$

$$\left.\left(\frac{\delta_{\sum_s x_{bs}} - R_b \delta_{\sum_s y_{bs}}}{E\sum_s y_{bs}}\right)\right]$$

$$= \frac{1}{E \sum_s y_{as} \; E \sum_s y_{bs}} \left[E \, \delta_{\sum_s x_{as}} \, \delta_{\sum_s x_{bs}} \right.$$

$$- R_a \; E \, \delta_{\sum_s y_{as}} \, \delta_{\sum_s x_{bs}} - R_b \; E \, \delta_{\sum_s x_{as}} \, \delta_{\sum_s y_{bs}}$$

$$\left. + R_a \, R_b \; E \, \delta_{\sum_s y_{as}} \, \delta_{\sum_s y_{bs}} \right]$$

$$= \frac{1}{E \sum_s y_{as} \; E \sum_s y_{bs}} \left[\text{Cov} \left(\sum_s x_{as}, \sum_s x_{bs} \right) \right.$$

$$- R_a \; \text{Cov} \left(\sum_s x_{bs}, \sum_s y_{as} \right) - R_b \; \text{Cov} \left(\sum_s x_{as}, \sum_s y_{bs} \right)$$

$$\left. + R_a \, R_b \; \text{Cov} \left(\sum_s y_{as}, \sum_s y_{bs} \right) \right]$$

Using Theorem 1 and Corollary 1.1 and rearranging terms,

$$\text{Cov} \left(\frac{\sum_s x_{as}}{\sum_s y_{as}}, \frac{\sum_s x_{bs}}{\sum_s y_{bs}} \right)$$

$$= E \sum_s \left[\left(\frac{E \sum_s x_{as}}{E \sum_s y_{as}} \right) \left\{ \frac{x_{as1} - x_{as2}}{E \sum_s x_{as}} - \frac{y_{as1} - y_{as2}}{E \sum_s y_{as}} \right\} \right.$$

$$\left. \left(\frac{E \sum_s x_{bs}}{E \sum_s y_{bs}} \right) \left\{ \frac{x_{bs1} - x_{bs2}}{E \sum_s x_{bs}} - \frac{y_{bs1} - y_{bs2}}{E \sum_s y_{bs}} \right\} \right]$$

Yields

$$\text{Var}(x) = \text{Var}\, x_a + \text{Var}\, x_b + 2 \, \text{Cov}(x_a, x_b)$$

$$= P_a^2 \; E \sum_s \left[\left(\frac{E \sum_s x_{as}}{E \sum_s y_{as}} \right) \left\{ \frac{x_{as1} - x_{as2}}{E \sum_s x_{as}} - \frac{y_{as1} - y_{as2}}{E \sum_s y_{as}} \right\} \right]^2$$

$$+ P_b^2 \; E \sum_s \left[\left(\frac{E \sum_s x_{bs}}{E \sum_s y_{bs}} \right) \left\{ \frac{x_{bs1} - x_{bs2}}{E \sum_s x_{bs}} - \frac{y_{bs1} - y_{bs2}}{E \sum_s y_{bs}} \right\} \right]^2$$

$$+ 2 P_a P_b \; E \sum_s \left[\left(\frac{E \sum_s x_{as}}{E \sum_s y_{as}} \right) \left\{ \frac{x_{as1} - x_{as2}}{E \sum_s x_{as}} \right. \right.$$

$$\left. - \frac{y_{as1} - y_{as2}}{E \sum_s y_{as}} \right\} \left(\frac{E \sum_s x_{bs}}{E \sum_s y_{bs}} \right) \left\{ \frac{x_{bs1} - x_{bs2}}{E \sum_s x_{bs}} \right.$$

$$\left. \left. - \frac{y_{bs1} - y_{bs2}}{E \sum_s y_{bs}} \right\} \right]$$

$$= E \sum_s \left[\sum_{a=a}^{b} \left(P_a \left(\frac{E \sum_s x_{as}}{E \sum_s y_{as}} \right) \right. \right.$$

$$\left. \left. \left\{ \frac{x_{as1} - x_{as2}}{E \sum_s x_{as}} - \frac{y_{as1} - y_{as2}}{E \sum_s y_{as}} \right\} \right) \right]^2$$

This can be directly generalized to the case where there are k different classes.

———○○○———

APPENDIX II

VARIANCE OF X''

The estimator is

(1)
$$x'' = \sum_a \frac{\sum_c \frac{\sum_s (x'_{acs1} + x'_{acs2})}{\sum_s (Z'_{cs1} + Z'_{cs2})} Z_c}{\sum_c \frac{\sum_s (y'_{acs1} + y'_{acs2})}{\sum_s (Z'_{cs1} + Z'_{cs2})} Z_c} y_a$$

$$= \sum_a \frac{x'_a}{y'_a} y_a$$

The variance of x'' is

(2) $\text{Var}(x'')$

$$= \sum_a \text{var}\left(\frac{x'_a}{y'_a} y_a\right) + \sum_a \neq \sum_b \text{Cov}\left(\frac{x'_a}{y'_a} y_a, \frac{x'_b}{y'_b} y_b\right)$$

(3) $$\text{Var}\left(\frac{x'_a}{y'_a} y_a\right)$$

$$= y_a^2 \left\{\left[\text{Var}(x'_a) + \left(\frac{Ex'_a}{Ey'_a}\right)^2 \text{Var}(y'_a) - 2\left(\frac{Ex'_a}{Ey'_a}\right)\text{Cov}(x'_a, y'_a)\right] \Big/ \left[E(y'_a)\right]^2\right\}$$

$$= \frac{y_a^2}{(Ey'_a)^2}\left[\text{Var}\left(\sum_c \frac{\sum_s (x'_{acs1} + x'_{acs2})}{\sum_s (Z'_{cs1} + Z'_{cs2})} Z_c\right) + \left(\frac{Ex'_a}{Ey'_a}\right)^2 \text{Var}\left(\sum_c \frac{\sum_s (y'_{acs1} + y'_{acs2})}{\sum_s (Z'_{cs1} + Z'_{cs2})} Z_c\right) - 2\left(\frac{Ex'_a}{Ey'_a}\right)\text{Cov}\left(x'_a, y'_a\right)\right]$$

By using Theorem 4 where $\sum_s x'_{acs} = (x'_{acs1} + x_{acs2})$, variance of $\left(\frac{x'_a}{y'_a} y_a\right)$ becomes

$$\text{Var}\left(\frac{x'_a}{y'_a} y_a\right)$$

$$= \frac{y_a^2}{(Ey'_a)^2}\left\{E \sum_s \left[\sum_c Z_c \left(\frac{E\sum_s x'_{acs}}{E\sum_s y'_{acs}}\right)\left\{\frac{x'_{acs1} - x'_{acs2}}{E \sum_s x'_{acs}} - \frac{Z'_{cs1} - Z'_{cs2}}{E \sum_s Z'_{cs}}\right\}\right]^2 + \left(\frac{Ex'_a}{Ey'_a}\right)^2 E \sum_s \left[\sum_c Z_c \left(\frac{E \sum_s y'_{acs}}{E \sum_s Z'_{as}}\right)\left\{\frac{y'_{acs1} - y'_{acs2}}{E \sum_s y'_{acs}} - \frac{Z'_{cs1} - Z'_{cs2}}{E \sum_s Z'_{cs}}\right\}\right]^2$$

$$- 2\left(\frac{Ex'_a}{Ey'_a}\right) E \sum_s \left[\sum_c Z_c^2 \left(\frac{E \sum_s x'_{acs}}{E \sum_s Z'_{cs}}\right)\right.$$

$$\left\{\frac{x'_{acs1} - x'_{acs2}}{E \sum_s x'_{acs}} - \frac{Z'_{cs1} - Z'_{cs2}}{E \sum_s Z'_{cs}}\right\}\left(\frac{E \sum_s y'_{acs}}{E \sum_s Z'_{cs}}\right)$$

$$\left.\left\{\frac{y'_{acs1} - y'_{acs2}}{E \sum_s y'_{acs}} - \frac{Z'_{cs1} - Z'_{cs2}}{E \sum_s Z'_{cs}}\right\}\right]\Bigg\}$$

$$= \frac{y_a^2}{(E y'_a)^2} E \sum_s \left[\sum_c Z_c \left\{\left(\frac{E \sum_s x'_{acs}}{E \sum_s Z'_{cs}}\right)\right.\right.$$

$$\left(\frac{x'_{acs1} - x'_{acs2}}{E \sum_s x'_{acs}} - \frac{Z'_{cs1} - Z'_{cs2}}{E \sum_s Z'_{cs}}\right) - \left(\frac{E x'_a}{E y'_a}\right)$$

$$\left.\left.\left(\frac{E \sum_s y'_{acs}}{E \sum_s Z'_{cs}}\right)\left(\frac{y'_{acs1} - y'_{acs2}}{E \sum_s y'_{acs}} - \frac{Z'_{cs1} - Z'_{cs2}}{E \sum_s Z'_{cs}}\right)\right\}\right]^2$$

Using Part 2 of the proof of Theorem 4, (4) can be stated

$$(4)\quad \mathrm{Cov}\left(\frac{x'_a}{y'_a} y_a, \frac{x'_b}{y'_b} y_b\right) = y_a y_b \mathrm{Cov}\left(\frac{x'_a}{y'_a}, \frac{x'_b}{y'_b}\right)$$

The following covariances are proved in Part 3 of the proof of Theorem 4.

$$(5)\quad \mathrm{Cov}\left(\frac{x'_a}{y'_a} \frac{x'_b}{y'_b}\right)$$

$$= \frac{1}{E y'_a E y'_b}\left[\mathrm{Cov}\left(\sum_c \frac{\sum_s x'_{acs}}{\sum_s Z'_{cs}} Z_c, \sum_c \frac{\sum_s x'_{bcs}}{\sum_s Z'_{cs}} Z_c\right)\right.$$

$$- \frac{E x'_a}{E y'_a} \mathrm{Cov}\left(\sum_c \frac{\sum_s y'_{acs}}{\sum_s Z'_{cs}} Z_c, \sum_c \frac{\sum_s x'_{bcs}}{\sum_s Z'_{cs}} Z_c\right)$$

$$- \frac{E x'_b}{E y'_b} \mathrm{Cov}\left(\sum_c \frac{\sum_s x'_{acs}}{\sum_s Z'_{cs}} Z_c, \sum_c \frac{\sum_s y'_{bcs}}{\sum_s Z'_{cs}} Z_c\right)$$

$$\left. + \frac{E x'_a E x'_b}{E y'_a E y'_b} \mathrm{Cov}\left(\sum_s \frac{\sum_s y'_{acs}}{\sum_s Z'_{cs}} Z_c, \sum_s \frac{\sum_s y'_{bcs}}{\sum_s Z'_{cs}} Z_c\right)\right]$$

$$(6)\quad \mathrm{Cov}\left(\sum_c \frac{\sum_s x'_{acs}}{\sum_s Z'_{cs}} Z_c, \sum_c \frac{\sum_s x'_{bcs}}{\sum_s Z'_{cs}} Z_c\right)$$

$$= \sum_c Z_c^2 \mathrm{Cov}\left(\frac{\sum_s x'_{acs}}{\sum_s Z'_{cs}}, \frac{\sum_s x'_{bcs}}{\sum_s Z'_{cs}}\right)$$

$$= \sum_c Z_c^2 \left[E \sum_s \left\{\left(\frac{E \sum_s x'_{acs}}{E \sum_s Z'_{cs}}\right)\right.\right.$$

$$\left(\frac{x'_{acs1} - x'_{acs2}}{E \sum_s x'_{acs}} - \frac{Z'_{cs1} - Z'_{cs2}}{E \sum_s Z'_{cs}}\right)$$

$$\left.\left.\left(\frac{E \sum_s x'_{bcs}}{E \sum_s Z'_{cs}}\right)\left(\frac{x'_{bcs1} - x'_{bcs2}}{E \sum_s x'_{bcs}} - \frac{Z'_{cs1} - Z'_{cs2}}{E \sum_s Z'_{cs}}\right)\right\}\right]$$

Writing the other three covariance terms similarly, where $\Delta x'_{acs} = x'_{acs1} - x'_{acs2}$

$$(7)\quad \mathrm{Cov}\left(\sum_c \frac{\sum_s y'_{acs}}{\sum_s Z'_{cs}} Z_c, \sum_c \frac{\sum_s x'_{bcs}}{\sum_s Z'_{cs}} Z_c\right)$$

$$= \sum_c Z_c^2 \left[E \sum_s \left\{\left(\frac{E \sum_s y'_{acs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{acs}}{E \sum_s y'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right.\right.$$

$$\left.\left.\left(\frac{E \sum_s x'_{bcs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta x'_{bcs}}{E \sum_s x'_{bcs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right\}\right]$$

$$(8)\quad \mathrm{Cov}\left(\sum_c \frac{\sum_s x'_{acs}}{\sum_s Z'_{cs}} Z_c, \sum_c \frac{\sum_s y'_{bcs}}{\sum_s Z'_{cs}} Z_c\right)$$

$$= \sum_c Z_c^2 \left[E \sum_s \left\{\left(\frac{E \sum_s x'_{acs}}{E \sum_s Z'_{cs}}\right)\right.\right.$$

$$\left(\frac{\Delta x'_{a.cs}}{E \sum_s x'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\left(\frac{E \sum_s y'_{bcs}}{E \sum_s Z'_{cs}}\right)$$

$$\left.\left.\left(\frac{\Delta y'_{bcs}}{E \sum_s y'_{bcs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right\}\right]$$

$$(9)\quad \mathrm{Cov}\left(\sum_c \frac{\sum_s y'_{acs}}{\sum_s Z'_{cs}} Z_c, \sum_c \frac{\sum_s y'_{bcs}}{\sum_s Z'_{cs}} Z_c\right)$$

$$= \sum_c Z_c^2 \left[E \sum_s \left\{\left(\frac{E \sum_s y'_{acs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{acs}}{E \sum_s y'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right.\right.$$

$$\left.\left.\left(\frac{E \sum_s y'_{bcs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{bcs}}{E \sum_s y'_{bcs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right\}\right]$$

Substituting equations (6), (7), (8), and (9) into equation (5),

$$(10)\ \mathrm{Cov}\left(\frac{x'_a}{y'_a}, \frac{x'_b}{y'_b}\right) = \frac{1}{E y'_a \, E y'_b}\left[E \sum_s \sum_c Z_c^2\right.$$

$$\left\{\left[\left(\frac{E \sum_s x'_{acs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta x'_{acs}}{E \sum_s x'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right.\right.$$

$$- \left(\frac{E x'_a}{E y'_a}\right)\left(\frac{E \sum_s y'_{acs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{acs}}{E \sum_s y'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\Big]$$

$$\left[\left(\frac{E \sum_s x'_{bcs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta x'_{bcs}}{E \sum_s x'_{bcs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right) - \left(\frac{E x'_b}{E y'_b}\right)\right.$$

$$\left.\left.\left.\left(\frac{E \sum_s y'_{bcs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{bcs}}{E \sum_s y'_{bcs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right]\right\}\right]$$

Substituting equations (3) and (10) into equation (2),

Var (x'')

$$= \sum_a \frac{y_a^2}{(E y'_a)^2} E \sum_s \left[\sum_c Z_c \left\{\left(\frac{E \sum_s x'_{acs}}{E \sum_s Z'_{cs}}\right)\right.\right.$$

$$\left(\frac{\Delta x'_{acs}}{E \sum_s x'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right) - \left(\frac{E x'_a}{E y'_a}\right)\left(\frac{E \sum_s y'_{acs}}{E \sum_s Z'_{cs}}\right)$$

$$\left.\left.\left(\frac{\Delta y'_{acs}}{E \sum_s y'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right\}\right]^2 + \sum_{a \neq b} \sum \frac{y_a \, y_b}{E y'_a \, E y'_b}$$

$$E \sum_s \left[\sum_c Z_c^2 \left\{\left[\left(\frac{E \sum_s x'_{acs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta x'_{acs}}{E \sum_s x'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right.\right.\right.$$

$$- \left(\frac{E x'_a}{E y'_a}\right)\left(\frac{E \sum_s y'_{acs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{acs}}{E \sum_s y'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\Big]$$

$$\left[\left(\frac{E \sum_s x'_{bcs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta x'_{bcs}}{E \sum_s x'_{bcs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right) - \left(\frac{E x'_b}{E y'_b}\right)\right.$$

$$\left.\left.\left.\left(\frac{E \sum_s y'_{bcs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{bcs}}{E \sum_s y'_{bcs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right]\right\}\right]$$

$$= E \sum_s \left[\sum_a \frac{y_a}{E y'_a} \sum_c Z_c \left\{\left(\frac{E \sum_s x'_{acs}}{E \sum_s Z'_{cs}}\right)\right.\right.$$

$$\left(\frac{\Delta x'_{acs}}{E \sum_s x'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right) - \left(\frac{E x'_a}{E y'_a}\right)$$

$$\left.\left.\left(\frac{E \sum_s y'_{acs}}{E \sum_s Z'_{cs}}\right)\left(\frac{\Delta y'_{acs}}{E \sum_s y'_{acs}} - \frac{\Delta Z'_{cs}}{E \sum_s Z'_{cs}}\right)\right\}\right]^2$$

APPENDIX III

SOME PROGRAMMING NOTES

Types of problems that arise when trying to implement the Keyfitz variance estimator are discussed in the section Health Interview Survey Variances. These problems are of a theoretical nature. In addition to these there are difficulties in programming the formulae for the SR PSU's and NSR PSU's.

In the Health Interview Survey, approximately 115,000 persons are interviewed each year. All the information gathered in these interviews undergoes the editing and processing outlined under Types of Statistics (see page 9). Neither the basic weights nor any of the intermediate weights, i.e. nonresponse adjusted weight, are retained. Also the formation is dispersed onto four different magnetic tapes. For illustrative purposes, consider the statistic the total number of acute conditions in the United States for fiscal year 1964. For this one statistic the following estimates are needed (using data for the entire year):

1. Nonresponse adjusted number of acute conditions for each of the 242 non-self-representing PSU's and for each of the 2,320 segments in the self-representing PSU's
2. Nonresponse adjusted estimate of the population in 60 age-sex-color classes for each of the 242 NSR PSU's and 2,320 segments
3. The simple inflated estimate of the population in 24 color-residence-region classes for each of the 242 NSR PSU's
4. Number of acute conditions in each of the 60 age-sex-color classes after full inflation using total sample
5. 1964 population figure for each of the 60 age-sex-color classes
6. Number of acute conditions in each of the 24 color-residence-region classes after full inflation using just the NSR PSU's
7. Population in each of 720 age-sex-color-residence-region classes after full inflation using just the NSR PSU's
8. 1960 census figure for each of the 24 color-residence-region classes
9. A "P" factor for each of the 242 NSR PSU's

Besides the estimates the 242 NSR PSU's and 2,320 segments must be numbered and paired. All these pieces of information are required before any mathematical calculations can be done. In order to get these estimates and figures, an intermediate weight is required—the nonresponse adjusted weight, data from different tapes, and outside information.

A brief outline follows of the programming necessary to accomplish the above objectives.

The first major step was to break the variance formulas into pieces that would help in programming The variance formulas (5) and (7) are

$$\hat{\text{Var}}\ (x''_{SR}) = \sum_{s} \left\{ (x'_{s1} - x'_{s2}) - \sum_{a} \frac{x''_a}{y_a} (y'_{as1} - y'_{as2}) \right\}^2$$

and

$$\hat{\text{Var}}\ (x''_{NSR}) = \sum_s \left\{ (2\,P_{s2}\ x'_{s1} - 2\,P_{s1}\ x'_{s2}) \right.$$

$$- \sum_a \frac{x''_a}{y_a}\ (2\,P_{s2}\ y'_{as1} - 2\,P_{s1}\ y'_{as2})$$

$$\left. - \sum_c \frac{2\,P_{s2} Z'_{cs1} - 2\,P_{s1}\ Z'_{cs2}}{Z_c} \left(x''_c - \sum_a \frac{x''_a}{y_a}\ y''_{ac} \right) \right\}^2$$

Rewriting the formulas

$$\hat{\text{Var}}\ (x''_{SR}) = \sum_s \left\{ \left(x'_{s1} - \sum_a \frac{x''_a}{y_a}\ y'_{as1} \right) \right.$$

$$\left. - \left(x'_{s2} - \sum_a \frac{x''_a}{y_a}\ y'_{as2} \right) \right\}^2$$

$$= \sum_s \left\{ (x''_{SR\text{-}s1}) - (x''_{SR\text{-}s2}) \right\}^2$$

$$\text{Var}\ (x''_{NSR})$$

$$= \sum_s \left\{ 2\,P_{s2} \left(x'_{s1} - \sum_a \frac{x''_a}{y_a}\ y'_{as1} - \sum_c \frac{Z'_{cs1}}{Z_c} \right. \right.$$

$$\left. (x''_c - \sum_a \frac{x''_a}{y_a}\ y''_{ac}) \right) - 2\,P_{s1} \left(x'_{s2} - \sum_a \frac{x''_a}{y_a}\ y'_{as2} \right.$$

$$\left. \left. - \sum_c \frac{Z'_{cs2}}{Z_c}\ (x''_c - \sum_a \frac{x''_a}{y_a}\ y''_{ac}) \right) \right\}^2$$

$$= 4 \sum_s \left\{ P_{s2} \left(x'_{s1} - \sum_a \frac{x''_a}{y_a}\ y'_{as1} - \sum_c \frac{Z'_{cs1}}{Z_c}\ x''_c \right. \right.$$

$$\left. + \sum_c \frac{Z'_{cs1}}{Z_c} \sum_a \frac{x''_a}{y_a}\ y''_{ac} \right) - P_{s1} \left(x'_{s2} - \sum_a \frac{x''_a}{y_a}\ y'_{as2} \right.$$

$$\left. \left. - \sum_c \frac{Z'_{cs2}}{Z_c}\ x''_c + \sum_c \frac{Z'_{cs2}}{Z_c} \sum_a \frac{x''_a}{y_a}\ y''_{ac} \right) \right\}^2$$

$$= 4 \sum_s \left\{ P_{s2}\ x''_{NSR\text{-}s1} - P_{s1}\ x''_{NSR\text{-}s2} \right\}^2$$

The formulas were separated into the estimates needed for each NSR PSU or SR segment for data processing purposes. Viewing the problem in this way facilitated programming.

A package of seven programs was written. The purpose, input, and output of the individual programs are as follows (see figure I):

1. Input—HIS tapes having the final weight for each interviewed person

   Purpose—To remove the first and second stage ratio adjustment factors from each person's final weight in order to get back to the nonresponse adjusted weight

   Output—Tapes having nonresponse adjusted weight for each person

2. Input—Tapes from run 1

   Purpose—To compute the statistics x'_{si} and the population estimates y'_{asi}

   Output—Tapes are sorted by PSU. Within each PSU the statistics x'_{si} and population estimates y'_{asi} are in order

3. Input—HIS regular tapes from which the estimates x are computed

   Purpose—To compute the x''_c's

   Output—Tapes are sorted by region, residence, and race

4. Input—HIS regular tapes

   Purpose—To compute x''_a's

   Output—Tapes are sorted by sex, age, and color

5. Input—The y'_{asi} computed in run 3, tapes from run 4, and y_a values from census

   Purpose—To calculate the quantity $x'_{si} - \sum_a \frac{x''_a}{y_a} y'_{asi}$ for each NSR PSU and SR segment

   Output—Tapes having $x'_{si} - \sum_a \frac{x''_a}{y_a} y'_{asi}$

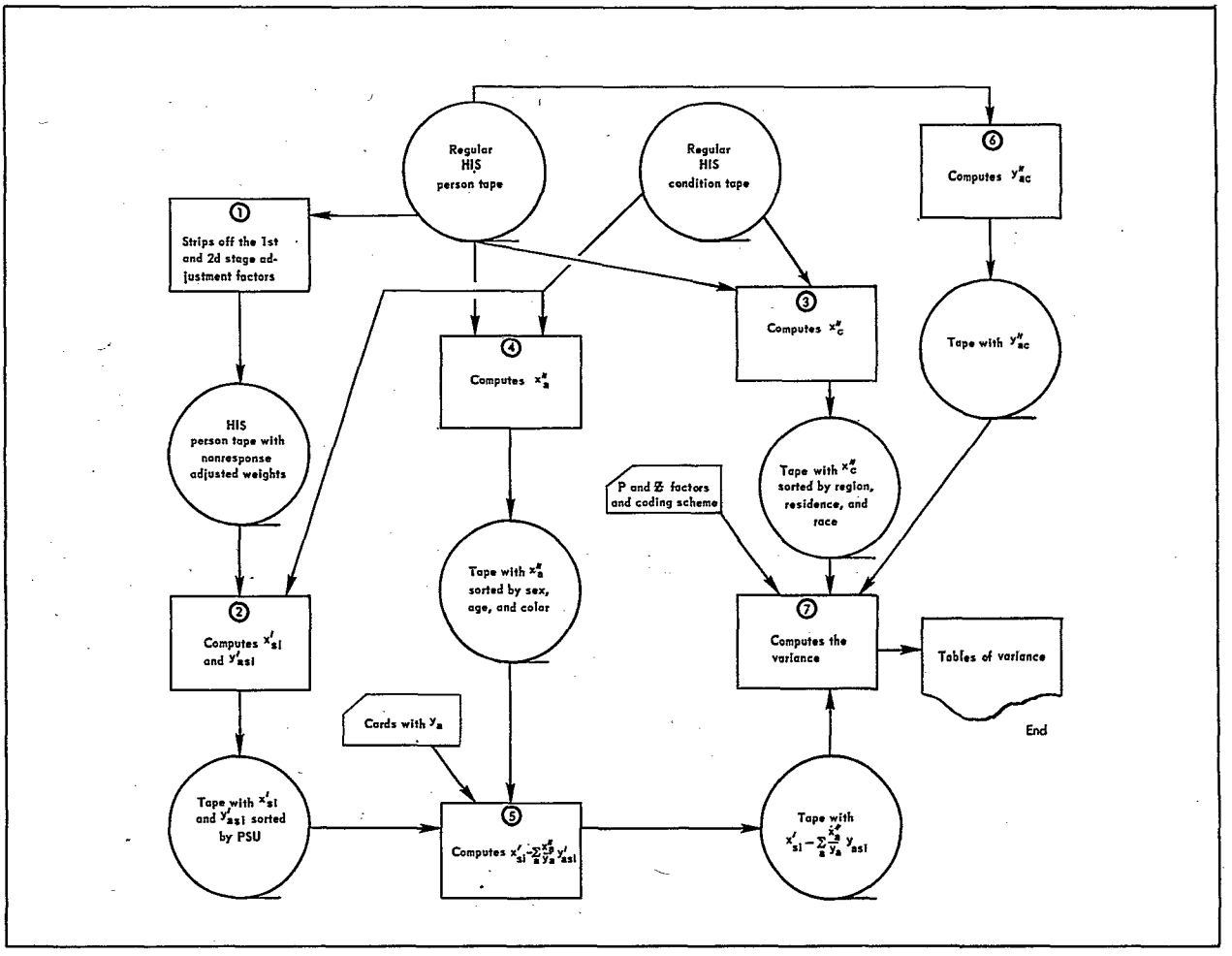

Figure I. Flowchart of the package of computer programs.

6. Input—HIS regular tapes

   Purpose—To calculate population estimates y''_{ac}

   Output—Tapes with the values

7. Input—Tapes from runs 3, 5, and 6, P and Z factors, and the coding scheme for matching and pairing PSU's and segments

   Purpose—To pair PSU's and segments, to match P and Z factors with appropriate PSU's, and to do all the rest of the calculations

   Output—The estimates x″ and their variances

☆ U. S. GOVERNMENT PRINTING OFFICE — 1970 — 396-245 P. O. 47

VITAL AND HEALTH STATISTICS PUBLICATION SERIES

Formerly Public Health Service Publication 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.

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.

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.

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.

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.

Series 11. Data from the Health Examination Survey.—Data from direct examination, testing, and measurement of national samples of the civilian, noninstitutional population provide the basis for two types of reports: (1) estimates of the medically defined prevalence of specific diseases in the United States and the distributions of the population with respect to physical, physiological, and psychological characteristics; and (2) analysis of relationships among the various measurements without reference to an explicit finite universe of persons.

Series 12. Data from the Institutional Population Surveys.—Statistics relating to the health characteristics of persons in institutions, and their medical, nursing, and personal care received, based on national samples of establishments providing these services and samples of the residents or patients.

Series 13. Data from the Hospital Discharge Survey.—Statistics relating to discharged patients in short-stay hospitals, based on a sample of patient records in a national sample of hospitals.

Series 14. Data on health resources: manpower and facilities.—Statistics on the numbers, geographic distribution, and characteristics of health resources including physicians, dentists, nurses, other health occupations, hospitals, nursing homes, and outpatient facilities.

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

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

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, hospital experience in the last year of life, medical care during pregnancy, health insurance coverage, etc.

For a list of titles of reports published in these series, write to:

Office of Information
National Center for Health Statistics
Public Health Service, HSMHA
Rockville, Md. 20852

DHEW Publication No. (HRA) 74-1288
Series 2-No. 38

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE
Health Resources Administration
5600 Fishers Lane
Rockville, Maryland 20852

OFFICIAL BUSINESS
Penalty for Private Use $300

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF HEW

HEW 390

THIRD CLASS
BLK. RT.