DATA EVALUATION AND METHODS RESEARCH

Series 2
Number 52

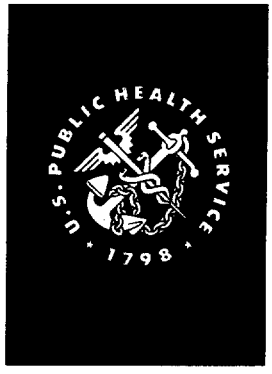
Reliability of Estimates With Alternative Cluster Sizes in the Health Interview Survey

Evaluation of loss in precision due to clustering of households.

DHEW Publication No. (HSM) 73-1326

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

Health Services and Mental Health Administration
National Center for Health Statistics
Rockville, Md. April 1973



Vital and Health Statistics-Series 2-No. 52

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402
Price 40 cents domestic postpaid or 30 cents GPO Bookstore

NATIONAL CENTER FOR HEALTH STATISTICS

OFFICE OF STATISTICAL METHODS

Vital and Health Statistics-Series 2-N0. 52

DHEW Publication No. (HSM) 73-1326
Library of Congress Catalog Card Number 72-600131

FOREWORD

This report presents the findings of one of a series of research projects designed to evaluate the efficiency of the sample design of the Health Interview Survey. The study focuses on only one feature of the design—namely, cluster size and associated intraclass correlations. Although valuable insights for designing samples can be obtained from measures of intraclass correlations, the optimum cluster size cannot be determined without knowledge of costs associated with the cluster sizes. At the present time, a detailed unit cost study is being conducted for the Center by the U.S. Bureau of the Census as part of the on-going Health Interview Survey. Nevertheless, it was decided to publish this report before the cost data became available since the estimates of intraclass correlations would be useful in the interim for designing State and local health surveys.

The study was conducted by the U.S. Bureau of the Census under a contractual arrangement with the Center. Theoretical and operational aspects of the study were performed by members of the Statistical Methods Division of the Census Bureau with principal responsibilities being shared by Joseph Waksberg, Robert H. Hanson, and Curtis A. Jacobs, who also prepared the text of this report.

For all studies conducted under contract, the NCHS staff develops general specifications for the study and works with the contractor on methodology and on technical decisions during the course of the study. These responsibilities were discharged by E. Earl Bryant.

MONROE G. SIRKEN

CONTENTS

RELIABILITY OF ESTIMATES WITH ALTERNATIVE CLUSTER SIZES IN THE HEALTH INTERVIEW SURVEY

Joseph Waksberg, Robert H. Hanson, and Curtis A. Jacobs,
Statistical Methods Division, U.S. Bureau of the Census,
Social and Economic Statistics Administration

INTRODUCTION

Objectives

The practice of grouping the elements of the universe of study into clusters and sampling the clusters is a common feature of sample designs. Dealing with clusters of elements almost invariably increases the sampling error of the statistics estimated from the sample. However, clustering usually reduces the per-unit costs of the sampling and data collection. As the cluster size is increased, the costs of conducting a survey are usually reduced but the sampling errors of the statistics generally increase for a given sample size. The increase in variance due to clustering may be considered as a "loss" in the reliability of a statistic. The purpose of this study is to obtain measurements of the loss in reliability due to clustering of households using various cluster sizes for data collected by the Health Interview Survey (HIS).

The HIS Sample

The HIS program began in July 1957. The design of this sample has had some modification since that time, but the essential features remain the same. The original design of the HIS is described in the Public Health Service publication No. 584–A2, "The Statistical Design of the Health Interview Survey." [1] The following summarizes the features of the design that are relevant to this study.

The HIS sample is a stratified multi-stage cluster design of approximately 44,000 designated occupied housing units selected for interview over the course of 1 year to provide estimates of selected health characteristics of the U.S. civilian noninstitutional population. The first stage of the design consists of a selection of primary sampling units (PSU's), one from each of 357 strata. The strata were constructed using geographic and demographic characteristics of the PSU's. A PSU is typically a county, a group of contiguous counties, or a standard metropolitan statistical area (SMSA). One hundred and twelve of the strata contain only one PSU and are called self-representing (SR); these PSU's are chiefly the larger SMSA's and contain about 56 percent of the U.S. population as of the 1960 Census. Each of the remaining 245 strata contain two or more PSU's—one of which has been selected as the sample PSU. These PSU's are termed nonself-representing (NSR) PSU's. The selected NSR PSU's represent themselves and the other PSU's in their strata. The sample PSU's were selected from the strata with probability proportional to the 1960 Census population of the PSU.

The second stage of sampling consists of a systematic selection of clusters of housing units within each sample PSU. The set of housing units in a sample cluster which are designated for interview in a given year is called a segment. Since July 1962, the HIS has made use of two types of segments which are referred to as list segments and area segments. One type of list segment used to represent units existing at the time of the 1960 Census is called a *B* segment. These segments are selected from the units recorded in

[1] U.S. Department of Health, Education and Welfare: *The Statistical Design of the Health Household Interview Survey*, PHS Pub. No. 584–A2, Series H–2. Public Health Service, Washington, D.C., U.S. Government Printing Office, July 1958.

the 1960 Census Listing Books and are made up of an average size of nine units (a more complete description of the *B* segment is given below). This segment type is used predominantly in urban or urban-type communities; about two-thirds of the total sample is comprised of such list sample segments. A second type of list segment is termed a *P* segment. These are clusters of units for which building permits have been issued since the 1960 Census and are used in the same areas as *B* segments. The remainder of the sample, mostly in the rural portions of the country, is comprised of area sample segments. Area sample segments are defined as small geographic areas constructed to have an expected size of nine 1960 Census housing units. Area segments and *P* segments were not used in this study for reasons mentioned later. Prior to July 1962, all segments used in the HIS were of the area sample type.

The segments comprising the sample of 44,000 designated housing units are randomly allocated to each of the 52 weeks of 1 year by procedures which equalize the proportion of sample cases in geographic subdivisions (central city, balance of urbanized areas, balance of urban areas, and rural areas) so that each weekly sample is a representative subsample of the total. Each year a new sample of units is used and a unit in the sample is interviewed only once.

The *B* segment as originally established for the HIS was usually composed of nine housing units selected systematically at the rate of one in two from a cluster of 18 units recorded on successive lines in the 1960 Census Listing Book. Thus, for areas covered by *B* segments, the HIS sample was made up of "noncompact" segments of nine housing units selected out of the cluster of 18 units. A segment of nine housing units chosen from nine consecutive lines in the 1960 Census Listing Book would be called a "compact" segment.

Noncompact segments are used in the HIS because their variances per unit in sample are presumed to be generally smaller than for compact segments. Although the travel costs associated with noncompact segments are somewhat higher than for compact segments, the lower variance for noncompact segments has been assumed to produce a more efficient sample. Theoretical considerations and empirical studies made for other surveys have suggested the validity of these assumptions. However, there is a scarcity of real data on the relationships which would permit selection of the optimum segment size for the HIS on a purely objective basis.

This study is designed to provide the data needed to compare the variances of samples for various types and sizes of segments, e.g., compact segments of two households, noncompact segments of nine households selected out of clusters of 18, compact segments of six. These data are needed to determine an optimum segment type for HIS, but information on unit costs is also necessary to examine alternative clustering methods. Data on costs are not considered in this report. In addition, no attempt was made to measure what effect the intraclass correlation has on the total variance of the HIS sample because of the restrictions of the data necessary to get the correlations. These restrictions are listed under the heading "Assumptions" in the next section.

STUDY METHODOLOGY

Measurements

An expression showing the effect of different size clusters on the relvariance of a sample estimate when random sampling of clusters and of listing units within sample clusters is used, is:[2]

$$V^2(\bar{x}) \doteq \frac{V^2}{m\bar{n}}[1+(\bar{n}-1)\,\delta_c] \qquad (1)$$

Equation (1) assumes the finite correction factor is insignificant and also assumes no variability in segment size. V^2 is the population relvariance between listing units, m is the number of segments in the sample, δ_c is the intraclass correlation between listing units randomly selected within clusters of size c, and $\bar{n}$ is the number of sample listing units per segment. We assume $\bar{n}$ is constant for all segments. The total number of listing units in the sample is given by $m\bar{n}$.

The intraclass correlation is a "measure of homogeneity" within clusters. When the listing units are relatively homogeneous within the clusters, i.e., when they are very similar with respect to some characteristic, the δ_c between listing units within clusters for the characteristic, will be high positive. Conversely, if the listing units within clusters are relatively heterogeneous with respect to the characteristic, the δ_c will be low positive or in unusual situations, even negative. In a survey using households as clusters of persons, the intraclass

[2] Hansen, M., Hurwitz W., and Madow, W.: *Sample Survey Method and Theory*, Vol. I and Vol. II, New York: John Wiley & Sons, Inc., 1953. (Vol. I., p. 259, equation (8.1).)

correlation among persons within households for a characteristic such as race would be high positive, as all members of households are typically of the same race. For the characteristic "males, age 25–44" the intraclass correlation would be very low positive or even negative as there is usually only one such person in a household.

The intraclass correlation has a range between $-\frac{1}{\bar{N}-1}$ and 1, where $\bar{N}$ is the number of sample and nonsample listing units within a cluster. Clusters of moderately large size may be efficient sampling units when the δ_c between listing units within clusters is a low positive number or negative, and will be less efficient when the δ_c is a high positive number. If δ_c is near zero, the characteristic is distributed throughout the population essentially as a random variable.

Since the expression $\frac{V^2}{m\bar{n}}$ is the relvariance for a simple random sample of $m\bar{n}$ units selected with replacement, the term $(\bar{n}-1)\ \delta_c$ of equation (1) is a measure of the loss in relvariance due to clustering. Usually, as the size of the cluster increases, the cluster becomes more heterogeneous and δ_c decreases. Thus, it is expected that δ_c for segments of nine compact units would be larger than the δ_c for 18 compact units, i.e.

$$\delta_{18} < \delta_9$$

for positive intraclass correlations. Thus,

$$1+(9-1)\ \delta_{18} < 1+(9-1)\ \delta_9. \qquad (2)$$

Expression (2) shows the relationship of the loss factors between using segments of nine noncompact units sampled out of clusters of 18 and using compact segments of nine. However, it has been observed that the decrease in δ_c as the size of a cluster increases is relatively slight, so that with compact clusters, the δ_c will usually satisfy the inequality:

$$1+(9-1)\ \delta_9 < 1+(18-1)\ \delta_{18}.$$

Shortly after the 1960 Census, it was decided to use noncompact segments of nine units selected out of clusters of 18 units for a substantial part of the HIS sample. This choice of segment is based on the relationship given in expression (2). Other factors, including knowledge of travel costs at that time, suggested this segment size.

A subsequent review of the cost figures in 1967 suggested that segments of six (rather than the nine then in use) could be employed with only a slight increase in cost. Since the expression

$$1+(6-1)\ \delta_{18} < 1+(9-1)\ \delta_{18} \qquad (3)$$

will be true for positive δ, the segment size was changed in July 1968 to noncompact segments of six units selected out of clusters of 18.

For random samples of clusters of equal size, the ratio of relvariances of alternative designs in analytic form can be expressed as functions of the parameters δ_c and $\bar{n}$. Although equations (1), (2), and (3) are appropriate for simple random sampling of clusters and may not strictly apply to systematic samples such as are used in the HIS, empirical studies have shown that the actual relationships are quite similar. Therefore, the relvariance of the different designs for this study have been expressed in the form of

$$1+(\bar{n}-1)\ \hat{\delta}_c.$$

The $\hat{\delta}_c$ is not necessarily exactly the same as the intraclass correlation which arises from the use of simple random samples; but, it is a useful way of summarizing the results and permitting inferences for different types of sample designs. Because HIS is a systematic sample, the expression $\hat{\delta}_c$ is defined to be the intraclass correlation in the loss function when dealing with a systematic sample of compact clusters containing c units.

Assumptions

To produce useful results for this study, the following assumptions and restrictions are necessary:

1. Only B segments which had at least nine designated housing units in sample were used for this study. If more than nine units were designated, only the first nine units were used in order to eliminate variation in the size of the segments. Those B segments having fewer than nine units were omitted; relatively few segments were deleted for this reason. Also, a few segments were systematically deleted to equalize the number per week in the sample. This was done to satisfy a basic assumption of the method of estimating the variance which requires the expected values of each member of a paired sample to be equal.

2. Area segments were not included because there is substantial variability in the size of these segments and the consideration relating to the selection of housing units within segments (i.e., nine units out of clusters of 18 units) does not apply in the same way for area segments.

3. Data for segments in self-representing PSU's were used. The segments in nonself-representing PSU's were eliminated for two reasons:

a. A large proportion of the samples from these PSU's are made up of area segments.

b. The between primary-sampling-unit component of the variance is eliminated by dealing only with self-representing PSU's and the estimates of $\hat{\delta}_c$ then reflect only the loss in variance due to clustering of housing units within segments.

4. It was assumed that data collected from pairs of adjacent weeks had the same expected values. Since the sample selected for each week is assumed to be an independent sample of the total population, the variance for a characteristic can be estimated by the sum of squares of the differences between pairs of weeks.[3] The equation used is:

$$s^2(x, \bar{n}) = \sum_{i=1}^{T} [x(i,1) - x(i,2)]^2 \qquad (4)$$

where $x(i, j)$ is the estimate of the characteristic x for the j^{th} week in the i^{th} pair of weeks ($j = 1, 2$) and $T = 78$. T is the number of pairs of weeks for the 3-year period of this study. Equation (4) represents the variance of the estimator

$$x' = \sum_{i=1}^{78} \sum_{j=1}^{2} x(i, j).$$

The relvariance of x' is obtained by dividing equation (4) by $(x')^2$, the square of the estimate.

5. The variances or relvariances of two simple random samples of the same population are inversely proportional to the two sample sizes. The same relationship is assumed to be true for the systematic sample used for the HIS; that is, if the size of an unclustered systematic sample is doubled, the variance is assumed to be halved.

Estimating the Loss Factor

Equation (4) was used to calculate estimates of variances and relvariances for the following set of sample designs:

(a) A systematic unclustered sample of m housing units. An unclustered sample has one housing unit per "segment." This relvariance appears in equation (1), in the right member as $\frac{V^2}{m\bar{n}}$ with $\bar{n} = 1$.

(b) A systematic sample of m clusters of six housing units with a noncompact segment of three housing units selected from each cluster, i.e., equation (1) with $\bar{n} = 3$, $c = 6$.

(c) A systematic sample of m clusters of 12 housing units with a noncompact segment of six housing units selected from each cluster, i.e., $\bar{n} = 6$, $c = 12$.

(d) A systematic sample of m clusters of 18 housing units with a noncompact segment of nine housing units selected from each cluster, i.e., $\bar{n} = 9$, $c = 18$.

The relvariances computed in (b), (c), and (d) appear in equation (1) as $V^2(\bar{x})$ for the appropriate segment sizes.

Using the assumption that the relvariance of a systematic sample is inversely proportional to the sample size, the relvariances computed from (a) were adjusted so the sample sizes were consistent with the relvariances of the designs described in (b), (c), and (d).

As suggested by equation (1), the relvariances computed for (b), (c), and (d) were divided by the adjusted estimates of relvariance of the unclustered design. These ratios represent measures of the factors which occur as a result of clustering, i.e., $[1 + (\bar{n} - 1)\,\hat{\delta}_c]$.

In order to generalize the results to alternative sample designs using other combinations of $\bar{n}$ and c, we assumed that for systematic samples, as for unrestricted random samples, the loss factor can be expressed by the quantity $1 + (\bar{n} - 1)\,\hat{\delta}_c$. Since the $\bar{n}$ for the three clustered designs listed above are known ($\bar{n} = 3$, 6, or 9), estimates of the values of $\hat{\delta}_c$ can be obtained.

Several special features were introduced in the calculation of the relvariances of the sample designs (a), (b), (c), and (d):

1. To estimate the relvariance of the unclustered design, the sums $x(i, 1)$ and $x(i, 2)$ were computed for the weeks in each pair (i) using one housing unit per segment. As each of the segments used contained nine housing units, nine independent estimates of variance without clustering were computed. The nine estimates were aggregated to give an estimate of the variance of an unclustered systematic sample and divided by the square of the sum of the nine estimates of the characteristic to give the relvariance.

2. Similarly, to produce estimates of relvariance for a clustered design of compact segments of six housing units, an independent estimate of the variance for a noncompact segment with $\bar{n} = 3$ was computed using the first set of three designated housing units in each segment within the given

[3] Keyfitz, N.: Sampling theory where two units are selected from each stratum, *JASA*, Vol. 52, Dec. 1957, p. 503. (For proof of this estimator, see appendix I.)

week. A second estimate employed the second set of three housing units in each segment and a third estimate used the third set of three housing units. This gave three estimates of the variance using noncompact segments of three selected out of clusters of six units, i.e., ($\bar{n}=3$, $c=6$). These three estimates were aggregated and divided by the square of the sum of the three estimates of the characteristic to give the relvariance.

3. Unweighted data were used to avoid the problem of interpreting the effect of the varying weights which are applied during the several stages of HIS estimation.

4. The listing unit for this study is a housing unit from which a response is expected. Nonresponses occur because people refuse to be interviewed, no one can be found at home, and because of vacant or demolished houses. Such nonresponses are included among the listing units to determine the segment size at the time of selection and the position of the interviewed housing units within the segment. The nonresponses are assigned a response of zero. A zero response for a demolition, a vacant, or a noninterview is consistent with the estimation process used. The refusal and no-one-at-home noninterviews are represented in the actual survey by modifying the weights of interviewed households but, in this study, a zero response has been assigned to them since an adjustment of the weights of the interviewed households would not reflect the physical location of the nonresponses within the segment. Approximately 44,000 occupied housing units are designated for interview in the sample for a typical fiscal year. Of this number, about 2,000 occupied units are visited but interviews are not obtained because the occupants are not found at home after repeated calls or are unavailable for some other reason. In addition to the 44,000 there are also about 9,900 sample units which are visited but are found to be vacant or otherwise not to be enumerated. About 19,660 housing units are in the sample for this study during each of the 3 years. There are, on the average, 378 housing units in the sample each week consisting of 42 segments of nine units.

5. The items were tabulated using 3 years of sample data comprising a total of about 59,000 designated housing units in list sample segments in self-representing PSU's. The 3 years of data are from fiscal years 1964, 1965, and 1966. The data of the 1965 and 1966 fiscal years are from the same set of clusters, but from the two different half-samples (segments) within the clusters. The housing units used in fiscal year 1965 are the first systematic half-samples of housing units within the set of clusters, and the fiscal year 1966 housing units are the second systematic half-sample. The two half-samples are mutually exclusive. The data for fiscal year 1964 is the second systematic half-sample selected out of a set of clusters of 18 units adjacent to the clusters used for fiscal years 1965 and 1966.

Reliability of Estimates of Loss

Estimates of the coefficients of variation of the intraclass correlations have been made for selected items; generalizations have been made to other items, and to the losses in variance. The estimator for the variance—equation (4)—uses the square of the differences between two adjacent weeks; each week is considered an independent estimate of the 2-week period. To determine the variance of the intraclass correlations, a similar estimator is used assuming the expected values of the variances for two adjacent pairs of weeks are equal and independent. The number of segments for each week were made equal within a consecutive 4-week period defined as strata (by random deletion of segments as necessary) to make the assumption as realistic as possible.

The estimated reliability[4] of the losses in variance due to clustering can be generalized as follows:

For loss factors due to clustering which imply a $\hat{\delta}_c$ of .3 or larger, the coefficient of variation (CV) in the loss due to clustering for segments of size $\bar{n} = 9$, $c = 18$ is approximately 5 percent and the coefficient of variation for the estimated $\hat{\delta}_c$ is around 7 percent. These and other findings are summarized in table A.

TABLE A. *Reliability of estimated loss factors and intraclass correlations for noncompact segments of nine housing units*

Estimated loss factor $1+(9-1)\hat{\delta}_{18}$	Estimated values of $\hat{\delta}_{18}$	Coefficient of variation of	
		$[1+(9-1)\hat{\delta}_{18}]$	$(\hat{\delta}_{18})$
		Percent	*Percent*
3.4 and over	.3 or larger	5	7
3.4 to 2.6	.3 to .2	5 to 10	7 to 16
2.6 to 1.8	.2 to .1	10 to 15	16 to 22
1.8 to 1.48	.1 to .06		22 to 45
1.48 to 1.24	.06 to .03	15 to 20	45 to 100
1.24 to 1.16	.03 to .02	More than 20	More than 100
Less than 1.16	Less than .02	Undefined	Undefined

[4]For computation procedure, see appendix II.

The standard error of the estimate is given by the product of the estimate and its coefficient of variation; thus, if $\hat{\delta}_{18}=.1$ and its coefficient of variation is 22 percent, its standard error is

$$.1 \times .22 = .022.$$

For most values of $\hat{\delta}_c$ estimated in this report, the standard errors are so large that tests of differences between the $\hat{\delta}_c$ for an item with different values of c will show the differences are not significant.

DESCRIPTION OF TABLES

Table 1

Table 1 presents for selected items collected by the HIS the loss factors due to clustering, the intraclass correlations, and the ratio of unclustered systematic sampling of housing units to simple random sampling of households. In addition, the number of sample cases having the characteristic in a typical year's sample is shown for each item along with the percentage of an appropriate total represented by these sample counts.

Columns (b)–(d) show the loss factors due to clustering, $[1+\hat{\delta}_c(\bar{n}-1)]$, for noncompact clusters of size $\bar{n}=3$, $c=6$, and $\bar{n}=6$, $c=12$ and $\bar{n}=9$, $c=18$.

Columns (e)–(g) present estimated intraclass correlations $\hat{\delta}_c$ for compact segments of size $c=6$, 12, and 18 housing units. The values of the intraclass correlations may be used to compute other values of $1+(\bar{n}-1)\hat{\delta}_c$ where $\bar{n} \leq c$.

Column (h) shows the approximate number of sample cases having the given characteristic in the study for 1 year of HIS data in the portion of the sample used in this study. However, the loss factors and the intraclass correlations were calculated using 3 years of HIS data.

Column (i) shows the proportion of the appropriate total represented by the number of cases having the characteristics.

Where applicable, column (j) presents for person or household characteristics, the ratio of the estimated variance of a systematic sample of households to the variance of a simple random sample of persons or households. This ratio is:

$$\frac{S^2sys}{S^2srs} = \frac{\sum_{i=1}^{T} [x(i,1)-x(i,2)]^2}{npq} \quad \text{where}$$

n is the total number of sample persons (or households, depending on the characteristic), and p the proportion of sample cases having the characteristic with $q=1-p$. The numerator of this ratio, S^2sys, is the estimated variance of an unweighted sample total calculated for the unclustered systematic sample of housing units given by equation (1) with $\bar{n}=1$. The denominator of the ratio, S^2srs, is the estimated variance of an unweighted sample total based on a simple random sample of persons or households (depending on the characteristic).

This ratio demonstrates the effect on the variance of some, but not all, stages of sampling for the HIS design. A ratio which demonstrated the net effect of all stages of sampling for the HIS design would include variance components from the following sources: (for person items) clustering of persons within housing units, variation in size of housing units, clustering of housing units by segment, variation in number of housing units in the segment, the effect of systematic sampling within PSU's, the procedure used for the stratification and selection of PSU's, variation among PSU's within strata, and the estimation process. (Kish[5] defines such a ratio, the net effect of all these factors, as the "design effect.") Most of these components of variation contribute to the variance in a positive manner. However, stratification, ratio estimation, and systematic sampling usually reduce the overall variance. Thus one would expect the true HIS estimates to have a smaller variance than implied by the "design effect." However, due to the restrictions of this study, there is no empirical measurement of this hypothesis. The data for this report are based on a restricted subset of the segments in self-representing PSU's only. Consequently, the ratios $\frac{S^2sys}{S^2srs}$ in table 1 do not include the effect of all of the components above. For housing items, the ratios show the net effect of systematic sampling within self-representing PSU's only. For person items, the ratios also include the effect of within-household clustering of persons and variation in household size. Therefore, the factors for person items are expected to be larger than for housing items. The numerator also includes the effect of vacant housing units and noninterviews which are treated as zeros, the effect of these is not reflected in the denominator.[6] The result of this is to understate the gains expected from systematic sampling.

[5] Kish, L.: *Survey Sampling*, John Wiley & Sons, Inc., New York City, 1965.

[6] See "Housing and Person Items" for further discussion of the effect of noninterview.

TABLE 1. *Estimated loss factors and intraclass correlations for selected items*

Item	Loss factor $[1+(\bar{n}-1)\hat{\delta}_c]$[1]			Intraclass correlation[2] $\hat{\delta}_c$			Total sample cases per year[3] n	Percent of (indicated total)	S^2sys/S^2srs[4,7]
	$\bar{n}=3$ $c=6$	$\bar{n}=6$ $c=12$	$\bar{n}=9$ $c=18$	$c=6$	$c=9$	$c=18$			
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
DEMOGRAPHIC STATISTICS									
Person Items								(Interviewed persons)	
Total population	1.54	2.14	2.56	.270	.228	.196	54,700	100%	
Negro	1.98	3.44	4.78	.490	.487	.472	6,100	11	5.30
All other races	1.90	3.17	4.63	.449	.434	.453	810	1.5	5.01
Persons under 1 year of age	1.04	1.17	1.33	.020	.034	.041	1,100	2	1.07
Persons under 17 years of age	1.44	1.82	2.10	.222	.164	.137	19,200	35	3.66
Persons 17+ years of age	1.43	1.94	2.46	.216	.188	.182	35,500	65	1.92
Males 17+ years of age	1.16	1.42	1.76	.082	.084	.095	16,400	30	.80
Total males	1.36	1.78	2.08	.179	.156	.136	26,300	48	2.37
Total females	1.44	1.94	2.25	.220	.189	.156	28,400	52	2.11
Family income less than $2,000	1.20	1.42	1.73	.102	.085	.091	3,300	6	3.08
Family income $5,000 or more	1.47	2.05	2.44	.236	.211	.180	38,000	69	8.18
Household Items								(Interviewed HU's)	
Total interviewed households	1.46	1.86	2.31	.231	.172	.164	16,700	100%	
Negro and "other" race head of household	2.22	3.94	5.72	.612	.588	.591	1,900	11	.96
Household income less than $2,000	1.20	1.30	1.63	.098	.061	.079	1,600	10	.42
Household income $5,000 or more	1.48	2.10	2.60	.238	.221	.200	10,400	62	1.26
Labor Force								(Interviewed persons)	
Current activity—employed	1.13	1.45	1.98	.064	.089	.122	20,200	37%	1.21
Current activity—unemployed	1.06	1.09	1.32	.029	.019	.040	1,000	1.8	1.20
HEALTH STATISTICS									
Person Items								(Interviewed persons)	
Persons with 1+ condition	1.33	1.61	1.98	.165	.122	.122	27,800	51%	2.45
Males with 1+ condition	1.16	1.22	1.44	.082	.043	.056	12,700	23	1.24
Persons with 1+ chronic condition	1.30	1.53	1.88	.151	.105	.110	24,700	50	1.99
Males with 1+ chronic condition	1.12	1.14	1.32	.058	.028	.040	11,250	21	1.07
Persons with activity limitation (1, 2, 3)[5]	1.15	1.20	1.26	.073	.041	.033	5,400	10	1.34
Persons with activity limitation (1)[5]	1.06	1.11	1.23	.030	.022	.029	900	1.6	1.11
Condition Items								(Conditions)	
Number of chronic conditions for males	1.09	1.04	1.26	.045	.008	.032	23,400	41%	[7]
Number of chronic conditions for females	1.10	1.22	1.33	.049	.045	.041	28,600	51	[7]
Period Items								(Bed days)	
Number of bed days in last 2 weeks	1.04	1.19	1.08	.022	.038	.011	12,200	100%	[7]

See footnotes at end of table.

TABLE 1. *Estimated loss factors and intraclass correlations for selected items*—Con.

Item	Loss factor $[1+(\bar{n}-1)\hat{\delta}_c]$[1]			Intraclass correlation[2] $\hat{\delta}_c$			Total sample cases per year[3] n	Percent of (indicated total)	S^2_{sys}/S^2_{srs}[4]
	$\bar{n}=3$ $c=6$	$\bar{n}=6$ $c=12$	$\bar{n}=9$ $c=18$	$c=6$	$c=9$	$c=18$			
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
HEALTH STATISTICS—Con.								(Activity days)	
Restricted activity days in last 2 weeks	1.10	1.16	1.04	.051	.033	.005	32,000	100%	(7)
								(Days lost)	
Days lost from work in last 2 weeks	1.11	1.11	1.04	.056	.022	.000	7,600	100%	(7)
Smoking Items								(Interviewed persons)	
Persons 17+ years who now smoke more than 20 cigarettes per day[6]	1.01	1.23	1.45	.050	.045	.057	8,500	16%	1.27
Persons 17+ years whose length of time since last regularly smoked cigarettes is 6 mos.[6]	.99	.99	.93	−.003	−.003	−.009	400	7	.93
Persons 17+ who ever smoked 60 cigarettes a day during heaviest smoking[6]	.97	.96	.93	−.015	−.009	−.009	650	1.2	.98
CONDITION STATISTICS									
Number of Acute Conditions								(Acute conditions)	
Total persons	1.07	1.30	1.48	.035	.059	.060	4,600		
Negro persons	1.14	1.59	1.77	.070	.118	.096	440	9.5%	(7)
Males	.94	1.07	1.19	−.029	.014	.024	2,100	46	(7)
White females	1.09	1.14	1.29	.043	.028	.037	2,250	49	(7)
Persons with 1+ bed days	1.07	1.12	1.28	.037	.023	.035	2,100	46	(7)
Number of Acute Upper Respiratory Illnesses									
Total persons	.92	1.08	1.11	−.043	.015	.014	1,800	39%	(7)
Males	.91	1.01	1.12	−.044	.002	.014	781	17	(7)
Females 17–44 years	.99	1.16	1.18	−.004	.031	.022	350	7.6	(7)
Persons with 1+ bed days	.88	.92	.83	−.060	−.016	−.021	775	17	(7)
Number of Cases of Influenza								(Acute conditions)	
Total persons	.96	.96	1.06	−.018	−.008	−.008	680	15%	(7)
Males	.91	.90	.90	−.046	−.019	−.013	310	7	(7)
Negro males	.95	1.20	.98	−.025	.041	−.002	23	.5	(7)
Females	.96	.91	1.03	−.019	−.017	−.004	370	8	(7)
Persons with 1+ bed days	1.06	.97	1.33	.029	−.006	.027	490	10.5	(7)

See footnotes at end of table.

TABLE 1. *Estimated loss factors and intraclass correlations for selected items*—Con.

Item	Loss factor $[1+(\bar{n}-1)\hat{\delta}_c]^1$			Intraclass correlation[2] $\hat{\delta}_c$			Total sample cases per year[3] n	Percent of (indicated total)	S^2sys/S^2srs [4,7]
	$\bar{n}=3$ $c=6$	$\bar{n}=6$ $c=12$	$\bar{n}=9$ $c=18$	$c=6$	$c=9$	$c=18$			
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
CONDITION STATISTICS—Con.									
Number of Persons Injured								(Injured persons)	
Total persons	.92	.91	1.00	−.042	−.017	.000	600	100%	
Males	.81	1.19	.83	−.094	−.038	−.021	340	57	
Males 25–44 years	1.00	1.06	1.02	.000	.012	.002	75	12	
Females 25–44 years	1.02	1.22	1.23	.010	.045	.029	125	21	
Injuries by moving motor vehicle	1.01	.90	.86	.007	−.020	−.017	43	7	
Injuries at work, white persons	.96	1.07	1.17	−.018	−.014	.021	74	12	
Injuries at school	.98	.99	1.17	−.008	−.002	.021	35	6	
Injuries at home—inside and outside	.97	.92	.97	−.015	−.016	−.004	250	41	
Number of Persons with								(Interviewed persons)	
Tuberculosis	.87	.89	.86	−.066	−.022	−.017	140	26%	
Asthma	1.08	1.10	1.04	.040	.020	.004	1,510	2.8	
Diabetes	1.02	1.08	1.09	.010	.015	.011	675	1.2	
Diseases of the heart	.93	.82	.87	−.037	−.037	−.016	1,630	2.8	
Chronic bronchitis	1.06	1.21	1.19	.028	.041	.024	950	1.6	
Miscellaneous									
Number of restricted activity days for acute conditions (in past 2 weeks)	1.08	1.20	1.27	.042	.041	.034	11,267	(Activity days) 35%	
Number of restricted activity days for acute conditions (in past 2 weeks), males	.94	.91	.84	−.031	−.017	−.020	5,040	16	
Number of bed days for acute condition (in past 2 weeks)	1.04	1.13	1.19	.022	.026	.024	5,250	(Bed days) 43%	
Number of bed days for acute condition (in past 12 months)	1.18	1.16	1.16	.091	.032	.020	189,200		
Number of chronic conditions with bed days (in past 2 weeks)	1.10	1.30	1.26	.050	.059	.033	8,020	(Conditions) 15%	
Number of chronic conditions with bed days (in past 12 months)	1.11	1.15	1.17	.057	.030	.022	3,900	7.3	
Number of chronic conditions with activity limitation (1, 2, 3)	1.13	1.12	1.24	.063	.025	.030	16,800	32	
Number of chronic conditions for persons 45–64 years	1.004	1.11	1.20	.002	.022	.024	16,300	29	
Total conditions (acute plus chronic)	1.18	1.33	1.64	.091	.067	.080	56,600	100	

See footnotes at end of table.

TABLE 1. *Estimated loss factors and intraclass correlations for selected items*—Con.

Item	Loss factor $[1+(\bar{n}-1)\hat{\delta}_c]$[1]			Intraclass correlation[2] $\hat{\delta}_c$			Total sample cases per year[3] n	Percent of (indicated total)	S^2_{sys}/S^2_{srs}[4,7]
	$\bar{n}=3$ $c=6$	$\bar{n}=6$ $c=12$	$\bar{n}=9$ $c=18$	$c=6$	$c=9$	$c=18$			
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
HOSPITAL STATISTICS									
Number of Short-Stay Hospital Discharges in Past 6 Months for								(Total discharges)	
Total persons	1.002	1.11	1.26	.001	.022	.032	3,400	6%	
Persons 45–64 years	.99	1.13	1.47	−.003	.027	.059	595	1	
Persons under 14 years	.86	.85	.87	−.072	−.030	−.016	870	1.6	
Females	.95	.97	1.01	−.023	−.0054	.001	2,130	4	
Negro and all other races	1.19	1.44	1.69	.095	.089	.086	325	6	
All persons with family income under $4,000	1.02	1.20	1.18	.012	.041	.022	2,680	5	
Surgery	1.01	1.14	1.10	.033	.027	.025	2,070	4	
Hospital Days for All Short-Stay Discharges in Past 6 Months for								(Hospital days)	
Total persons	.93	.88	.85	−.033	−.024	−.018	33,900	100%	
Persons under 14 years	.99	1.17	1.37	−.004	.035	.046	3,550	10	
Persons 45–64 years	.96	1.02	1.10	−.023	.004	.013	11,590	34	
Females	.97	.80	.81	−.014	−.041	−.024	18,520	55	
Surgery	1.02	1.02	.82	.010	.004	−.023	16,640	49	
Number of Short-Stay Hospital Discharges in Past 12 Months for								(Total discharges)	
Total persons	1.11	1.35	1.62	.057	.069	.077	6,350	12%	
Persons 45–64 years	1.06	1.14	1.45	.033	.028	.056	1,110	2	
Persons under 14 years	.91	.94	.92	−.044	−.011	−.010	1,560	2.9	
Females	1.18	1.39	1.52	.089	.077	.065	4,040	7	
Negro and all other races	1.23	1.66	1.91	.114	.132	.114	610	1	
All persons with family income under $4,000	1.08	1.32	1.48	.039	.065	.060	4,980	9	
Surgery	1.12	1.34	1.54	.060	.068	.068	3,860	7	
Hospital Days for All Short-Stay Discharges in Past 12 Months for								(Hospital days)	
Total persons	.91	.91	.85	−.047	−.018	−.019	59,340	100%	
Persons under 14 years	.96	.99	1.20	−.018	−.002	.024	6,350	11	
Persons 45–64 years	.94	.99	1.08	−.029	−.002	.011	19,230	32	
Females	1.01	.92	.87	.002	−.016	−.016	33,370	56	
Surgery	.96	1.01	.90	−.020	.002	−.012	29,530	49	

[1] Losses in variance due to clustering for various values of $\bar{n}$ and c.

[2] Estimated intraclass correlations for compact clusters of c housing units.

[3] The approximate number of sample cases for 1 year of HIS data. The $\hat{\delta}_c$ intraclass correlations and variance loss factors were calculated using 3 years of HIS data.

[4] Ratio of the variance of an unclustered systematic sample of housing units to simple random sampling of persons (or housing units depending on the characteristics).

[5] Activity limitation is defined for persons with one or more chronic conditions as:

1. Cannot perform usual activity.
2. Can perform usual activity but limited in amount or kind.
3. Can perform usual activity but limited in outside activities.

[6] The variances and intraclass correlations for these items were calculated for 2 years of data only, FY 1965 and FY 1966.

[7] The value of $\frac{S^2_{sys}}{S^2_{srs}}$ is not shown because estimates of the term S^2_{srs} are not available.

Table 2

Table 2 shows the percent increase in variance over unclustered systematic sampling of housing units for compact and noncompact segments of size $\bar{n}=6$, and $\bar{n}=9$ for selected items. Columns (a) and (b) show the percent increase for compact segments of $\bar{n}=6$ and $\bar{n}=9$, respectively. The percent of increase in variance over unclustered systematic sampling for compact segments of $\bar{n}=9$ has been interpolated using the value of $\hat{\delta}_c$ between $\hat{\delta}_6$ and $\hat{\delta}_{12}$ as given in table 1.

Columns (c) and (d) give the increases in variance of the new HIS design ($\bar{n}=6$, $c=18$) and the old HIS design ($\bar{n}=9$, $c=18$), respectively.

Column (e) presents the percent increase in the variance of unclustered systematic sampling of housing units over the variance expected from simple random sampling of either persons or households, depending on the characteristics.

TABLE 2. *Percent increase in variance of systematic samples of clusters over unclustered systematic samples of housing units*

[Percent]

Characteristic	Compact segments		Noncompact segments		Systematic sampling of households to simple random sampling[3]
	Six units	Nine units	Six out of 18 units[1]	Nine out of 18 units[2]	
	(a)	(b)	(c)	(d)	(e)
Total population	135	199	98	156	...
Negro	245	391	236	378	430
All other races	225	360	227	363	401
Persons under 1 year	10	20	21	33	7
Persons under 17 years	111	161	69	110	266
Persons 17 years and over	108	156	91	146	92
Males 17 years and over	41	68	48	76	−20
Total males	90	134	68	108	137
Total females	110	161	78	125	111
Persons with family income less than $2,000	51	79	45	73	208
Persons with family income $5,000 or more	118	177	90	144	718
Total interviewed households	116	171	82	131	...
Negro and all other races heads of household	306	486	295	473	−4
Households with income less than $2,000	49	63	39	63	−58
Households with income $5,000 or more	119	182	100	160	26
Persons with current activity—employed	32	63	61	98	21
Persons with current activity—unemployed	15	21	20	32	20
Persons with 1+ condition	82	123	61	98	145
Males with 1+ condition	41	61	28	44	24
Persons with 1+ chronic condition	75	104	55	88	99
Males with 1+ chronic condition	29	35	20	32	7
Persons with activity limitation (1, 2, 3)	36	50	17	26	34

[1] Similar to HIS sample since July 1968.

[2] Similar to HIS sample prior to July 1968.

[3] Increase in variance of unclustered systematic sample of housing units over simple random sample of persons (or housing units depending on the characteristic).

... = Quantity not defined.

GENERAL OBSERVATIONS

Comments

The following comments on the results of the data indicate the restrictions one should be aware of in using the results of this report.

1. The loss factors due to clustering for all items in this study are based on simple inflation estimates which are unbiased if all assumptions are true. In using the results of this study for design purposes, one must remember that the estimation procedure for the survey uses ratio adjustments at several levels, particularly to known age-sex totals; also this study has produced variances under several restrictions of the data. Thus, even though the estimates of the intraclass correlations are valid, inferences as to the size of variances for the complete design are not valid for items which are strongly affected by ratio estimation, stratification, and choice of the PSU's used. The first nine items presented in table 1 are some of the control items used in deriving the estimates of data in the HIS survey. For these items, the variances of the HIS estimates (using all elements of the sample design and estimation procedure) are actually zero, regardless of the type and size of clusters used within PSU's.

2. Changing the size of the subsample drawn out of the cluster of 18 from $\bar{n}=9$, to $\bar{n}=6$ will reduce the variance of the estimate for items with a positive intraclass correlation for a constant sample size because the loss in variance is a function only of $\bar{n}$ (the subsample size) and not of the intraclass correlation.

3. All tests of significance of differences between any numbers used in this report have assumed independence between the two numbers forming the difference. This is a conservative approach since the correlation coefficient between the two numbers is usually positive. All tests were computed at the 95-percent confidence level.

4. Since the HIS is based on a sample of one-half of the housing units in each cluster, it is not possible to estimate directly the intraclass correlation for compact segments of $c=9$ or for any odd number c. Thus, it is not possible to show the corresponding differences in variances for compact and noncompact segments of nine for the current HIS design. Interpolation based on the values of $\hat{\delta}_c$ in table 1 may be used to estimate the value for a compact segment of nine units.

5. One should be aware that the effect on the variance contributed by the $\hat{\delta}_c$ is very small compared to the effect of $\bar{n}$ in the loss factor.

Housing and Person Items

The values of the intraclass correlations for person items will usually decrease as the cluster size increases. Table 1 shows that this appears to be the case for most items. However, the sampling errors of the values of $\hat{\delta}_c$ are so large,[7] one cannot make comparisons among individual values of $\hat{\delta}_c$ for a given item.

There is, in general, for person items evidence indicating noncompact segments have a lower variance than compact segments. One also expects the increase in variance over unclustered systematic sampling of housing units to be larger for compact segments than for noncompact segments; table 2 shows this is usually the case.

The ratio in column (e) of table 2 is expected to be negative for household items (i.e., a gain in efficiency using systematic sampling as opposed to simple random sampling). This expectation isn't necessarily the case but the apparent discrepancy can be explained by the following two reasons:

1. The numerator of the ratio in column (e) reflects the presence of noninterviews and vacant housing units, the effect of which is not present in the denominator. The implications of this are as follows: Approximately 0.2 of all designated units were vacant or noninterview. For random sampling, the relvariance based on all units can be expressed in terms of the relvariance of occupied units by the relationship: [8]

$$V^2 \text{ all} = \frac{V^2 \text{ (occupied)} + 0.2}{0.8} = 1.25\, V^2 \text{ (occupied)} + 0.25$$

If we assume the relationship for systematic sampling is similar, the ratio given in column (e) of table 2 can be expressed.

$$100\left[\frac{V^2 sys}{q/np} - 1\right] = 100\left[\frac{V^2 \text{ all}}{q/np} - 1\right] = 100\left[\frac{1.25 V^2_{occ}}{q/np} + \frac{0.25}{q/np} - 1\right]$$

[7] See "Reliability of Estimates of Loss" for coefficients of variation.

[8] Hansen, Hurwitz, and Madow: op. cit., Vol. I, page 604.

The first ratio in the brackets on the right would more nearly state the relationship of systematic sampling of housing units to simple random sampling. This suggests that the figures in column (e) of table 2 represent an understatement of the expected gains from systematic sampling.

2. Another reason for the small gains shown for systematic sampling may be due to the design of the HIS sample. The proportion of all clusters of nine units in the universe that are in the sample for the areas in this study in a given year is about 1 in 2,778. Thus, two clusters in the systematic sample are separated by about 25,000 housing units (about 80,000 persons). This sampling interval is so large that the expected effect of stratification in systematic sampling is not realized because the natural groupings in the population tend to make the "strata" more heterogeneous than would result for smaller sampling intervals.

Condition Items

The variances for compact and noncompact segments of size 6 differ only slightly for condition items. All of the observed differences are within sampling error. One expects the increase in variance over unclustered systematic sampling to be larger for compact than for noncompact segments, but the sampling errors of the $\hat{\delta}_c$'s for this study are so large that this hypothesis cannot be confirmed or refuted. In a large majority of the items dealing with acute conditions, upper respiratory illness, influenza, and persons injured, the intraclass correlation is estimated to be negative or a small positive number indicating that δ_c is probably very close to zero.

Hospital Items

Table 1 shows a large number of negative estimates of intraclass correlations for hospital and condition items. The sampling errors of these estimates are so large, however, that none of these negative values can be shown to be significantly different from zero; thus the conclusion that the $\hat{\delta}_c$'s are actually very nearly zero is made, suggesting the items are close to being randomly distributed in the population.

Conclusions

It is evident from the estimates of the intraclass correlations and the losses in variance given in tables 1 and 2 that for a majority of the hospital and condition items, the use of compact segments could be justified. Also the $\hat{\delta}_c$'s indicate the size of the cluster may be increased with minimal increase in the variance due to clustering. These conclusions do not apply to person or household items.

— ○○○ —

APPENDIX I

Estimator for Variance of Systematic Sample of Clusters

To estimate the variance of an estimator x', an annual characteristic from the HIS, consider a successive pair of 2 weeks as a "stratum of time." There are 26 such strata (pairs of weeks) in the course of a year. For a given stratum, each week in the pair is considered an independent sample for the same time period, so that

$$E[x(i,1)]=E[x(i,2)] \tag{1A}$$

where $x(i, j)$ is the estimate of a characteristic for the jth week in the ith pair of weeks ($j=1,2$).

The magnitude of each $x(i, j)$ is one-half the stratum total which is in turn equal to $1/26$ of the annual estimate. The problem is to find a variance estimator for the sum of the ith pair of weeks; i.e.

$$\text{Var}\,[x(i,1)+x(i,2)] \tag{2A}$$

and in summing the variances over all pairs of weeks (strata) to give an estimate of $\sigma_{x'}^2$ where x' is the estimated annual characteristic:

$$x'=\sum_{i=1}^{26}[x(i,1)+x(i,2)] \tag{3A}$$

Since $x(i, 1)$ and $x(i, 2)$ are independent random variables, it can be shown that in the ith stratum:

$$\text{Var}\,[x(i,1)+x(i,2)]=E[x(i,1)-x(i,2)]^2. \tag{4A}$$

Since the 26 strata are assumed to be independent

$$\text{Var}\left\{\sum_{i=1}^{26}[x(i,1)+x(i,2)]\right\}=E\left\{\sum_{i=1}^{26}[x(i,1)-x(i,2)]^2\right\}=\sum_{i=1}^{26}E[x(i,1)-x(i,2)]^2. \tag{5A}$$

The estimator is

$$\text{Var}\,[x(i,1)+x(i,2)]\doteq\sum_{i=1}^{26}[x(i,1)-x(i,2)]^2. \tag{6A}$$

The estimated variances resulting from the use of the estimator (6A) will depend in turn on the method of determining the estimates $x(i, j)$. However, as shown elsewhere, the variances of the estimates are influenced by the choice of the segment size $\bar{n}$. To indicate that the variances are related to the segment size, the estimator (6A) is written

$$s^2(x,\bar{n})=\sum_{i=1}^{26}[x(i,1)-x(i,2)]^2 \tag{4}$$

where it is understood that in (4), the estimates $x(i, 1)$ and $x(i, 2)$ are constructed using segments of $\bar{n}$ listing units.

———o o o———

APPENDIX II

Reliability of the Estimated Intraclass Correlation and Loss Factor

This appendix outlines the method used to calculate the relvariances of the loss factors and the relvariances of the intraclass correlations.

1. From appendix I equation (4), an estimate of the variance of a clustered systematic sample is:

$$s^2(x, \bar{n}) = \sum_{i=1}^{T} [x(i, 1) - x(i, 2)]^2 \qquad (1B)(4)$$

where $x(i,j)$ is the estimate of the characteristic for the jth week of the ith pair of weeks using segments of $\bar{n}$ listing units. When $\bar{n}=1$, expression (1B) is an estimate of the variance for an unclustered systematic sample; i.e., $x(i, j)$ is the estimate for the ijth week constructed from a systematic sample of segments of one listing unit.

The estimator (1B) assumes each of the 2 weekly estimates $x(i, j)$ in a successive pair of weeks is an independent estimate within the "stratum of time" represented by the pair of weeks. In similar manner, the relvariance of (1B) is estimated with the assumption each of the estimates $s^2(x, \bar{n})$ from two successive pairs of weeks is an independent estimate within a "stratum of time" represented by a set of 4 weeks.

2. To simplify the expressions for the estimates, the following notation is introduced. Let:

a) $u = s^2(x, \bar{n})$ the estimated variance of a clustered systematic sample and $t = s^2(x, 1)$, the estimated variance of an unclustered systematic sample.

b) $d(k, i)$, the squared difference between the estimates $x(i, j)$ based on clustered samples for the two members of the ith pair of weeks of the kth set of 4 weeks. $\left(i = 1, 2; k = 1, 2, \ldots, \frac{T}{2}\right)$. More specifically, within the kth set of 4 successive weeks, there are the following estimates for the 4 weeks based on clustered samples:

$$x(k, 1, 1)$$

$$x(k, 1, 2)$$

$$x(k, 2, 1)$$

$$x(k, 2, 2)$$

These terms are used to form the following two squared differences:

$$d(k, 1) = [x(k, 1, 1) - x(k, 1, 2)]^2 \qquad (2B)$$

$$d(k, 2) = [x(k, 2, 1) - x(k, 2, 2)]^2 \qquad (3B)$$

Using this notation, we could write equation (1B) as

$$s^2(x, \bar{n}) = \sum_{k=1}^{T/2} \sum_{i=1}^{2} d(k, i).$$

c) $\hat{d}(k, i)$ the squared difference between the estimates $x(i, j)$ based on unclustered samples $\left(i = \hat{d}1, 2; \quad k = 1, \ 2, \ldots, \ \frac{T}{2}\right)$. The expression $d(k, i)$ is adjusted to the sample size equivalent to $d(k, i)$. The equation (2B) and (3B) computed for segments of 9 for example, requires the accumulation of all 9 units in each segment for the estimate; however, nine estimates of $\hat{d}(k, i)$ for the unclustered case can be produced from this data by choosing one of the 9 units from each segment for each estimate.

3. With this notation, the ratio u/t is

$$u/t = 1 + \hat{\delta}_c (\bar{n} - 1) \qquad (4B)$$

With the usual Taylor approximation the relvariance of this ratio is

$$v^2(u/t) \doteq v^2(u) + v^2(t) - 2v(u, t) \qquad (5B)$$

a) The required relvariances are given by:

$$v^2(u) = \sum_{k=1}^{T/2} \frac{[d(k,1) - d(k,2)]^2}{[s^2(x,\bar{n})]^2} \qquad (6B)$$

$$v^2(t) = \sum_{k=1}^{T/2} \frac{[\hat{d}(k,1) - \hat{d}(k,2)]^2}{[s^2(x,1)]^2} \qquad (7B)$$

$$v(u,t) = \sum_{k=1}^{T/2} \frac{[d(k,1) - d(k,2)]\,[\hat{d}(k,1) - \hat{d}(k,2)]}{[s^2(x,\bar{n})]\,[s^2(x,1)]} \qquad (8B)$$

b) The coefficient of variation of the loss factor is given by the square root of equation (5B).

4. From equation (4B):

$$\hat{\delta}_c = \frac{1}{\bar{n}-1}\left[\frac{u}{t} - 1\right] \qquad (9B)$$

a) Using equation (5B), it follows that

$$v^2(\hat{\delta}_c) \doteq \frac{u^2}{(u-t)^2} v^2(u/t) \qquad (10B)$$

b) The coefficient of variation of $\hat{\delta}_c$ is given by the square root of equation (10B).

——— ○○○ ———

☆ U. S. GOVERNMENT PRINTING OFFICE : 1973 515-212/54

VITAL AND HEALTH STATISTICS PUBLICATION SERIES

Originally 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.

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