EMPIRICAL STUDY OF EFFECT ON VARIANCES DUE TO SAMPLING PAIRS OF CONSECUTIVE WEEKS IN A NATIONAL HEALTH INTERVIEW SURVEY EXPERIMENT

Iris Shimizu, National Center for Health Statistics

1. Introduction

The National Center for Health Statistics conducted an experiment in the National Health Interview Survey (NHIS) to measure the degree to which the use of proxy respondents affects NHIS statistics. For this experiment, the NHIS sample was divided during the second quarter of 1972 into two groups by taking pairs of consecutive weeks. Interviewers used the usual respondent rule, which allows proxy response, in one of these groups while in the second group they required all adults capable of being a self respondent to be one.

The experiment itself was described by Haase and Wilson (1972) while Kovar and Wright (1973), Sirken, Kovar, and Wright (1974) and White and Massey (1979) examined the effect which proxy response has on the NHIS estimates.

The purpose of this paper is to describe the methods and results of an investigation into the effects on sampling errors in the experiment caused by use of alternate pairs of weeks to form the control and experimental groups instead of alternate weeks or simple random sampling. The NHIS design is described in the next section while Section 3 presents the three experimental sampling designs considered in our study of affects on variation. The comparisons between the variances derived from the different experimental sampling designs are discussed in Section 4. Section 5 summarizes the findings.

2. Survey Design for NHIS

Data for the National Health Interview Survey are collected by personal interviews in a continuing nationwide sample of households. The survey collects information on personal and demographic characteristics, illness, impairments, chronic conditons, health resource utilization, and other health topics.

The NHIS uses a multistage probability design which permits a continuous sampling of the civilian noninstitutionalized population of the United States. The NHIS sample of households interviewed each week is representative of the target population and the weekly samples are additive over time.

The first stage of the sample design consists of about 375 primary sampling units (PSU's) selected from approximately 1,900 geographically defined PSU's. A PSU consists of a county, a group of contiguous counties or a standard metropolitan statistical area. Within PSU's, segments containing, on the average, four housing units each are selected.

The NHIS sample each year consists of approximately 12,000 segments, 50,000 sample housing units, and 120,000 persons.

Data collection for the survey is performed by the U.S. Bureau of the Census under specifications established by the NCHS in collaboration with the Census Bureau.

The NHIS weighting scheme consists of several factors: inflation by the reciprocal of the probability of selection, a nonresponse adjustment, a ratio adjustment to PSU decennial populations within 12 color-residence classes, and ratio adjustment to Census Bureau estimates of the U.S. totals within each of 60 age-sex-color cells.

3. Sampling Designs and Estimators

3.1. Alternate Pairs of Consecutive Weeks

In the NHIS respondent rule experiment, alternate pairs of consecutive weeks in the spring quarter (April 2 through June 30) of the ongoing NHIS were allocated to each of the two respondent rule groups. That is, if the weeks of the year are sequentially numbered from 1 to 52 starting with the week including January 1, the experiment began in week 14 and the "self respondent" rule was applied in the interviews conducted during weeks 15, 16, 19, 20, 23, and 24 (a total of 6 weeks). The usual respondent rule was applied to the interviews conducted during the remaining weeks of the spring quarter.

In order to formulate the estimators and variances for the statistics based on data collected from each of the two respondent groups, let:

- i denote individual week,
- j denote the pair of consecutive weeks where $i = 2j - 1$ and $i = 2j$,
- B be the set of alternate pairs of consecutive weeks included in either the experimental or the control sample,
- k_A = the number of pairs of consecutive weeks included in A,
- X = value for variable of interest,
- X_i' = annual estimate for X based on the i-th week of data, and
- $Y_j' = X_{2j-1}' + X_{2j}'$.

Then the annual estimator of X based on data collected in a sample of alternate pairs of consecutive weeks can be formulated as

$$\hat{X}_A = \frac{1}{2\,k_A} \sum_{j \in A} Y_j' . \qquad (3.1)$$

In order to compute the variances, it was assumed that the Y_j' are independent and identically distributed so the variance of (3.1) can be estimated by

$$\mathrm{Var}(\hat{X}_A) = \frac{1}{(2\,k_A)^2} \sum_{j \in A} \mathrm{Var}(Y_j')$$

$$= \mathrm{Var}(Y_j')/4\,k_A . \qquad (3.2)$$

We also assumed that the variance for samples of alternate unites can be approximated by the variance that would apply if a simple random of units were selected. That is,

$$\mathrm{Var}(Y_j') = \frac{\sum_{j\in A'} (Y_j')^2 - \left(\sum_{j\in A'} Y_j'\right)^2 / k_{A'}}{k_{A'} - 1} \quad (3.3)$$

where A' is a set of alternate pairs of weeks

3.2 Alternate Weeks Sampling Design

Other sampling designs could have been used for the NHIS experiment. Two of these other designs are considered in this investigation of effects on the resulting variances.

Alternate weeks instead of alternate pairs of consecutive weeks could have been used to select the two groups in the NHIS Study. Under this sampling design, let

B denote the set of alternate weeks included in either the control or the experimental sample,

k_B = number of weeks in set B.

The estimator could then be formulated as

$$\hat{X}_B = \frac{1}{k_B} \sum_{i\in B} X_i' \quad (3.4)$$

where X_i' is as defined earlier. If we assume that the weekly estimates are independent and identically distributed, the variance for (3.4) can be approximated by

$$\mathrm{Var}(\hat{X}_B) = \frac{1}{(k_B)^2} \sum_{i\in B'} \mathrm{Var}(X_i') = \mathrm{Var}(X_i')/k_B \ . \quad (3.5)$$

If we assume the simple random sample formula for variances may be used to approximate the variances from samples of alternate units, then

$$\mathrm{Var}(X_i') = \frac{\sum_{i\in B'} (X_i')^2 - \left(\sum_{i\in B'} X_i'\right)^2 / k_{B'}}{k_{B'} - 1} \quad (3.6)$$

where B' is a set of weeks which includes set B.

3.3 Simple Random Sampling Design

Another sampling design that could have been used in the NHIS experiment involves simple random samples of individuals in the target population. To compute the estimates under the assumption of simple random samples of individuals, NHIS sample frequency distribution data were used. To formulate the estimators used, let

g denote an individual,

C denote a set of individuals included in the simple random sample for either the experimental or the control group,

k_C = total number of individuals in set C,

H = total number of classes of people defined by some variable, e.g., the number of doctor visits, bed days,

m denote one of the H classes of people defined by some variable,

C_m denote the set of people in C who belong to class m,

f_m = number of people in C_m,

x_g = value of X for the g-th person, and

$\bar{x}_m = \sum_{g\in C_m} x_g / f_m$

is the sample average for the variable X among people in class m.

The simple random sample estimator for the mean of X (i.e., the sample mean) can then be formulated as

$$\hat{\bar{X}}_C = \sum_{m=1}^{H} \bar{x}_m f_m / k_C \quad (3.7)$$

and the corresponding variance estimator can be formulated as

$$\mathrm{Var}(\hat{\bar{X}}_C) = \frac{\sum_{m=1}^{H} (\bar{x}_m)^2 f_m - \left(\sum_{m=1}^{H} \bar{x}_m f_m\right)^2 / k_{C'}}{k_C (k_{C'} - 1)} \quad (3.8)$$

where set C' set of sampled individuals which includes set C.

For estimates of percents used in our investigation of effects, NHIS weekly estimates of percentages were used to produce estimates. To formulate the estimates based on this data, let

P = i-th week estimate of the percent P for the characteristic of interest,

D = set of weeks in which data were collected from the individuals in set C,

k_D = number of weeks in set D.

Then the estimates for percent P may be approximated by

$$\hat{P}_C \approx \bar{P}_D = \sum_{i\in D} P_i / k_D. \quad (3.9)$$

If $\hat{P}_C$ were actually based upon a simple random sample of individuals, then the corresponding variance may be estimated by

$$\mathrm{Var}(\hat{P}_C) = \hat{P}_C(1 - \hat{P}_C)/k_C \quad (3.10)$$

where k_C is defined as before.

4. Study Methods

The goal of this investigation is to empirically determine the increase, if any, in sampling errors that can result from the use of alternate pairs of consecutive weeks instead of alternate individual weeks or a simple random sample of individuals to form the groups in an NHIS experiment, and in particular the NHIS respondent rule experiment. For purposes of this study, it is assumed that the experimental and the control sample each include the same number of weeks. That is,

estimates for the usual rule group, as well as the self-respondent rule group, are assumed to bebased on six weeks of data. This keeps the variances in our investigation from being affected by sample size.

Also for the study, it is assumed that variances are unaffected by the respondent rule in affect at the time data are collected. This is indeed true if all data reported are correct and the response rates are unaffected by the respondent rule used in data collection.

The variables considered in this investigation are limited to those variables, or variables similar to those, used in the NHIS experiment for which both the NHIS sample frequency distribution and weekly estimates were available for some other data years. Data from other years are used to assure that the resulting sampling errors are independent of effects that may be caused by changing the respondent rule. The data needed from other years were available for only six of the variables eligible for consideration in this investigation. These six are displayed in the form of rates per 100 people in Table A.

The estimates and variances for the six variables are computed using the formulas in Section 3 and the NHIS data from the prior years. That is estimates and variances are computed assuming, in turn, each of the three sampling designs that could have been used in the NHIS experiment. Data from the 1969 NHIS is used in the study except for the simple random sample results for four of the study variables. The simple random sample results for these four are based on 1967 and 1971 since the frequency data required in the computations were not available for 1969. Under the assumption that the sampling errors for individual variables do not vary significantly from year to year under the same sampling design, the use of data from different years will yield the relationships among the sampling errors that would exist if all the sampling errors could be computed using data from a commmon year. Under the same assumption, the relationships observed in this investigation are considered independant of the data year.

While all variances are adjusted to reflect sample sizes equivalent to the number of sample cases in six weeks of the NHIS sample, 52 weeks of data are used in each of the variance computations. This is because the sample frequencies needed to compute estimates under the simple random sampling design were available for whole data years only. Use of data from comparable seasons is necessary to assure that seasonality in the data can not confound the differences between estimates and sampling errors for each parameter under the different sampling designs.

The relative standard errors (RSE's) which are the standard errors RSE's for the estimates divided by the estimates, themselves are computed to reduce sampling errors to percents of the corresponding estimates. The resulting RSE's are presented in Table A.

RSE's are compared in the following by use of ratios of the RSE's formed by dividing the RSE's from the alternate pairs of consecutive weeks sample by the corresponding RSE's from the other sample sampling designs. These ratios are also shown in Table A.

5. Findings

As expected, among the three sampling designs investigated in this study, the simple random sample design yielded the smallest RSE's for all of the variables included in the study and the sample of alternate pairs of consecutive weeks yielded the largest RSE's for most of the variables.

Comparing first the results from the alternate pairs of consecutive weeks sample with the those from the alternate weeks sample, it can be seen in column three of Table A that the ratios of RSE's for the last four variables are close to one. This implies for these four variables that pairing consecutive weeks to form the samples yields about the same sampling errors as does a sample of alternate weeks. However, for bed days and restricted activity days, pairing consecutive weeks increased the RSE's by 40 to 50 percent over those yielded by the alternate weeks sample. This implies the existance of correlation between estimates based upon data from consecutive weeks.This could be expected on the basis of procedures used to collect the data. Two week reference periods are used to collect data on bed days and restricted activity days. However, among the four variables whose RSE's appear to be about the same under the alternate pairs of consecutive weeks sample and the alternate weeks sample, only the variable "doctor visits" is collected with a reference period of two weeks. Reference periods of six or more months are used to collect the data for the other three variables.

The evidence in Table A indicates that a sample of alternate weeks yields overall smaller sampling errors than does a sample of alternate pairs of consecutive weeks. In the NHIS experiment which sparked this study, however, pairs of consecutive weeks were used instead of alternate weeks in order to avoid administrative difficulties. Among other things, it was thought that interviewers might get confused if the respondent rules were changed every week and thus add to the non-samplng errors. Hence, it is possible that even if alternate weeks were used there would not be any real gain.

Turning next to the comparison with simple random sample results, the interest is in the resulting design effects on the RSE's. The design effect on RSE's for a sampling design is the ratio of RSE's based on data collected under that sampling design divided by the corresponding RSE's based on a simple random sample. The ratios in column five of Table A are design effects on RSE's for the sample of alternate pairs of consecutive weeks. For the last four variables in the table, the design effects on the RSE's are about two, which is comparable to the design effects observed for NHIS estimates. However, the design effects on RSE's for bed days and restricted activity days are about double those of the other four variables. This again indicates that the sampling errors for estimates of bed days and restricted activity days are affected differently by the use of paired consecutive weeks than are those for the remaining variables. In particular, pairing consecutive weeks in the sample increased the sampling errors for bed days and restricted activity days relative to those based on simple random sample more than it did for the remaining variables.

6. Summary

We have empirically reviewed the effect on relative standard errors (RSE's) due to the use of alternate pairs of consecutive weeks instead of alternate weeks or simple random sampling to form the two groups in an NHIS experiment. On the basis of the limited study, it appears that pairing consecutive weeks in an NHIS experimental sample doubles or quadruples the sampling errors that would be obtained under simple random sampling. If a reference period of six or more months is used to collect the data for a variable, then pairing consecutive weeks yields sampling errors that approximate those that would be yielded by a sample of alternate weeks. However, if a reference period of two weeks is used to collect the data for a specific variable, then pairing consecutive weeks will probably yield larger sampling errors than would a sample of alternate weeks.

Thus, on the basis of this limited study, it appears that if one can eliminate potential increases in non-sampling error that may arise from frequent changes in data collection procdures or make those errors minimal in comparison to sampling errors, and if two week reference periods are used in the data collection, then pairs of consecutive weeks should not be used to form control and experimental groups in NHIS or NHIS type surveys.

REFERENCES

Haase, Kenneth W. and Wilson, Ronald W. (1972). "The Study Design of an Experiment to Measure the Effects of Using Proxy Responses in the National Health Interview Survey". Proceedings of the Social Statistics Section, American Statistical Association, Washington, D.C., pp. 289-293.

Kovar, M. G. and Wilson, R. W. (1976). "Perceived Health Status: How Good is Proxy Reporting?" Proceedings of the Social Statistics Section, Part II, American Statistical Association, Washington, D.C., pp. 495-500.

Kovar, M. G. and Wright, R. A. (1973). "An Experiment with Alternate Respondent Rules in the National Health Interview Survey." Proceedings of the Social Statistics Section, American Statistical Association, Washington, D.C., pp. 311-316.

Sirken, M. G., Kovar, M. G. and Wright, R. A. (1974, May 1). "Effects of Respondent Rules on the Estimates of the Health Interview Survey," National Center for Health Statistics unpublished draft report, Washington, D. C.

White, Andrew A. and Massey, James T. (1981). "Selective Reduction of Proxy Response Bias in a Household Interview Survey." Proceedings of the Social Statistics Section, American Statistical Association, Washington, D.C., pp. 211-216.

TABLE A: Relative Standard Errors for Selected Rates by Selected Sample Designs for NHIS Experiments and Their Ratios

Rates	Method of Sample Selection				
	Alternate Pairs of Consecutive Weeks	Alternate Weeks		Simple Random Sample	
	RSE (1)	RSE (2)	Ratio (3) (1)/(2)	RSE (4)	Ratio (5) (1)/(4)
Bed days per 100 persons	0.210	0.145	1.4	0.047	4.5
Restricted activity days per 100 persons	0.134	0.092	1.5	0.032	4.2
Percent of population having doctor visits in the last 6 months	0.016	0.014	1.1	0.007	2.3
Percent of population with limitation of activity due to chronic conditions	0.036	0.035	1.0	0.022	1.6
Hospital discharges per 100 persons	0.046	0.048	1.0	0.027	1.7
Doctor visits per 100 persons	0.056	0.050	1.1	0.025	2.3