EMPIRICAL DISTRIBUTIONS OF BALANCED HALF-SAMPLE REPLICATION VARIANCE ESTIMATOR AND LINEARIZATION VARIANCE ESTIMATOR

Judy A. Bean, University of Iowa

1. Introduction

The majority of sample surveys of any magnitude are complex multi-stage probability surveys. This type of survey may include 1) unequal probabilities of selection of the different sampling units in the population; 2) stratification of the units; 3) two or more stages of clustering; and 4) nonlinear estimation. The use of these complex surveys creates problems that cannot be solved by classical statistical theory. One of the problems is the estimation of the variance of the parameter estimator.

The solution three of the major survey organizations, the United States Bureau of the Census, the National Center for Health Statistics and Survey Research Center, University of Michigan, use is to calculate variance estimates from the sample data by one of the following general variance estimators: 1) balanced half-sample replication method; 2) linearization method; and 3) jackknife method. Because the mathematics is intractible, the properties of the estimators have not been derived.

Recently, Bean (1, 2) and Frankel (3) have investigated the estimators by Monte Carlo sampling from a completely specified universe. Simulation studies from an empirical distribution permit comparison of the methods under conditions of known population values, complete response from the sampling units, and no observational errors. In the present study, the empirical behavior of the balanced half-sample replication variance estimator and the linearization variance estimator was examined in terms of their histograms generated from Monte Carlo sampling of empirical data.

2. Methodology

2.1 Balanced Half-Sample Replication Method and Linearization Method.

Before describing the sample design of this study, the main features of the two variance estimators will be outlined.

Suppose we have a population of individuals grouped in clusters of households which are themselves grouped into larger clusters called primary sampling units (PSU's); then these PSU's are classified into L strata. Consider a survey whose design is two primary sampling units selected from each of the L strata with subsampling within each of the chosen PSU's. An estimate X' of the population parameter is computed from the X'_{hi}'s where X'_{hi} is the sample estimate for the i^{th} PSU in the h^{th} stratum.

To obtain an estimate of variance for X' by the replication procedure, a half-sample is created by randomly selecting one of the two PSU's from each of the L strata. Another estimate, X'_j, of the same population parameter is then made, utilizing only the data from the one half-sample. The quantity $(X'_j - X')^2$ is an estimate of variance of X'; however, this variance estimate has itself a high variance. To produce a less variable variance estimator, k such half-samples are drawn with the mean of the squared differences being the variance estimate.

$$\hat{\sigma}_{X'}^2 = \frac{1}{k} \sum_{j=1}^{k} (X'_j - X')^2$$

The linearization method is derived from the theorem by Keyfitz (4) which states that the variance of a sum of two independent estimates of a parameter equals the expected value of the square of the difference between them. This theorem can be extended to estimates produced when the design is two PSU's selected from each of the L strata, assuming selections among the strata are independent. The method consists of estimating the stratum totals and then linearly combining them to yield an estimate of variance.

$$X' = \sum_{h=1}^{L} (X'_{h1} + X'_{h2})$$

$$\hat{\sigma}_{X'}^2 = \sum_{h=1}^{L} (X'_{h1} - X'_{h2})^2$$

2.2 Universe.

The universe for the study consisted of morbidity data collected by the United States Health Survey (HIS) in 131,575 civilian noninstitutional individuals. A description of the survey has been published by the National Center for Health Statistics (6).

The sample design in HIS is a stratified two stage cluster sample with the first stage units (PSU's) being counties of the United States and the second units (ultimate sampling units) being clusters of 6 households within the counties. The 357 original PSU's in HIS were regrouped to form 148 PSU's; the ultimate sampling units containing less than 3 individuals were combined with similar units.

2.3 Sample Design.

The design used included unequal probabilities of selection, stratification and clustering. The 149 PSU's were classified into 19 strata, eight containing only one primary sampling unit. The first stage of sampling consisted of the independent selection of two PSU's drawn with replacement from eleven strata with probability proportional to size. The other eight strata entered the sample with a probability of one. In the second stage of sampling, the ultimate sampling units in selected PSU's were randomly subsampled with replacement.

In order to observe the behavior of the variance estimates for different sample sizes, three sample sizes were used (smaller = design I, intermediate = design II, largest = design III). Since the design called for the selection of two primary sampling units from 11 of the strata with the remaining 8 strata automatically being in the sample, the subsampling rate applied within each PSU was varied to achieve the different sample sizes. For each design 900 samples were independently drawn with a total of 2700 samples being selected.

2.4 Variables and Estimators.

In order to study the behavior of the variance estimates for different distributions, the following variables were selected: 1) family

income; 2) number of restricted activity days; 3) number of physician and dental visits; 4) number of days spent in short-stay hospitals; and 5) whether or not the person has seen a physician within a 12-month period. The average of variables 1 through 4 above per person per year were estimated as was the proportion of the population seeing a physician.

For each one of the 2700 samples drawn, a ratio post-stratified estimate of each of the five characteristics was computed. The post-stratification was performed for 24 ethnic-sex-age classes (white and non-white, male and female, ages 0-4, 5-14, 15-24, 25-44, 45-64 and 65+).

$$R' = \frac{\sum_{a=1}^{A} \frac{x'_a}{y'_a} y_a}{y}$$

where

R' = the final post-stratified ratio estimate of the x^{th} characteristic,

x'_a = the simple inflated estimate of the x^{th} characteristic for the a^{th} ethnic-sex-age class,

y'_a = the simple inflated estimate of the population in the a^{th} ethnic-sex-age class,

and

y_a = the known population in the a^{th} ethnic-sex-age class.

2.5 Estimators for the Variance of a Ratio Sample Statistic.

a. Replication Method. There are three different ways of estimating the variance of a statistic by using the balanced half-sample replication method; these are described in McCarthy (5). Variances were computed by all three versions but since one of them is the average of the other two, only it will be discussed. The variance estimator is

$$\hat{\sigma}^2(R') = \tfrac{1}{2}\left[\sum_{j=1}^{k} \frac{(R'_j - R')^2}{k} + \sum_{j=1}^{k} \frac{(R^*_j - R')^2}{k}\right]$$

where

k = the number of half-samples,

R' = the final post-stratified ratio estimate,

R'_j = the post-stratified ratio estimate secured from the j^{th} half-sample,

and

R^*_j = the post-stratified ratio estimate secured from the complement half-sample (the PSU's not in the half-sample).

The full orthogonal balance pattern was used to select the PSU's that form the repeated half and complement half-samples. The pattern is presented in Bean (2). For each R' computed, an estimate of variance was calculated using the above equation.

b. Linearization Method.

The estimator R' can be written as

$$R' = \frac{\sum_{h=1}^{L} (x'_{ah1} + x'_{ah2})}{\sum_{h=1}^{L} (y'_{ah1} + y'_{ah2})} y_a + \cdots$$

where

x'_{ahi} = the inflated estimate of the x^{th} characteristic for class a of the i^{th} PSU in the h^{th} stratum.

and

y'_{ahi} = the inflated estimate of the population in class a of the i^{th} PSU in the h^{th} stratum.

Then the linearization variance estimator of R' is

$$\hat{\sigma}^2(R') = \frac{\sum_{h=1}^{L}\left[(x'_{h1} - x'_{h2}) - \sum_{a=1}^{A} \frac{x''_a}{y_a}(y'_{ah1} - y'_{ah2})\right]^2}{y^2}$$

where

$$x''_a = \sum_{a=1}^{A} \frac{x'_a}{y'_a} y_a$$

For each of the 8 stratum consisting of only one PSU, two pseudo-PSU's were created from the sampled segments. The details are given in Bean (2). Again for each R' estimate, a variance estimate was produced from the above equation.

3. Summary of Empirical Work

Because of the volume of the primary estimates and variance estimates generated, only the results for the largest sample size, design III, are presented. In Table 1 can be seen the average of the 900 statistics calculated along with the mean variance estimate of the replication method and of the linearization method. The two means are very similar for all the variables studied.

Tables 2 and 3 present the frequency and the cumulative percent for the variables restricted activity days and proportion seeing a physician. The results are nearly the same.

Figures 1 through 5 show the histograms of the variance estimates produced by the two techniques. As can be seen, the distributions are very much alike between methods. Notice the distributions are skewed with bunching below the mean and a tail above; this skewness decreased as the sample size increased. Bean (2) compared the estimates in terms of their means, sampling variances and biases. Each estimator has a small bias.

Because the balanced half-sample replication method and the linearization method provide similar results, either may be used to calculate variance estimates for data gathered in a complex multistage probability sample survey. The validity of applying either variance estimator method to this type of survey has been reported on in detail elsewhere [Bean (1,2) and Frankel(3)]. The studies showed that the variance estimates produced by the methods are satisfactory.

ACKNOWLEDGMENT

The author wishes to thank the National Center for Health Statistics for providing the data and to thank Mr. Walt R. Simmons for help and encouragement received in doing the research.

REFERENCES

1. Bean, Judy A. "Behavior of Replication and Linearization Variance Estimators for Complex Multistage Probability Samples." University of Texas, 1973. (unpublished dissertation).

2. Bean, Judy A. "Distribution and Properties of Variance Estimators for Complex Multistage

Probability Samples." Vital and Health Statistics. (In press).

3. Frankel, Martin R. Inference to Survey Samples. An Empirical Investigation. Ann Arbor: University of Michigan, 1971.

4. Keyfitz, Nathan. "Estimates of Sampling Variance where Two Units are Selected from each Stratum." Journal of the American Statistical Association, 52, (1957), 503-510.

5. McCarthy, Philip J. "Replication: An Approach to the Analysis of Data from Sample Surveys." Vital and Health Statistics. PHS Publication No. 1000, Series 2, No. 14. Washington, D. C.: Government Printing Office, 1966.

6. National Health Survey. "The Statistical Design of the Health Household - Interview." Health Statistics. PHS Publication No. 584-A2. Washington, D. C.: Government Printing Office, 1958.

TABLE 1. A summary of the values obtained by repeated sampling for Design III.

Value	Family Income	Restricted Activity Days	Physician and Dental Visits	Short-Stay Hospital Days	Proportion Seeing Physician
Average Sample Estimate - R'	8392.80	14.6595	4.6548	1.0597	0.6840
Average Variance Estimate .. Replication	26554.80	0.9206	0.0178	7.0842×10^{-3}	8.3010×10^{-5}
Average Variance Estimate .. Linearization	26174.70	0.8915	0.0175	6.7380×10^{-3}	8.1135×10^{-5}

TABLE 2. The frequency and cumulative per cent of 900 variance estimates as calculated by the balanced half-sample replication variance estimator and by the linearization variance estimator for the variable average number of restricted activity days per person per year for Design III.

Variance Estimate	Replication Frequency	Replication Cumulative Per Cent	Linearization Frequency	Linearization Cumulative Per Cent
0.00-0.20	0	0.00	0	0.00
0.21-0.40	37	4.11	48	5.33
0.41-0.60	131	18.67	136	20.44
0.61-0.80	225	43.67	225	45.44
0.80-1.00	190	64.79	207	68.44
1.01-1.20	141	80.44	131	83.00
1.21-1.40	80	89.33	64	90.11
1.41-1.60	46	94.44	52	95.89
1.61-1.80	26	97.33	20	98.11
1.81-2.00	14	98.89	10	99.22
2.01-2.20	6	99.56	4	99.67
2.21-2.40	1	99.67	1	99.78
2.41-2.60	2	99.89	1	99.89
2.61-2.80	0	99.89	0	99.89
2.81-3.00	1	100.00	1	100.00

TABLE 3. The frequency and cumulative per cent of 900 variance estimates as calculated by the balanced half-sample replication variance estimator and by the linearization variance estimator for the variable proportion seeing a physician within the last 12 months for Design III.

Variance Estimate (var. times $x10^{-4}$)	Replication Frequency	Replication Cumulative Per Cent	Linearization Frequency	Linearization Cumulative Per Cent
0.00-0.10	0	0.00	0	0.00
0.11-0.20	1	0.11	2	0.22
0.21-0.30	9	1.11	10	1.33
0.31-0.40	35	5.00	43	6.11
0.41-0.50	75	13.33	79	14.89
0.51-0.60	110	25.56	114	27.56
0.61-0.70	141	41.22	147	43.89
0.71-0.80	123	54.89	124	57.67
0.81-0.90	99	65.89	93	68.00
0.91-1.00	69	73.56	69	75.67
1.01-1.10	73	81.67	53	81.56
1.11-1.20	42	86.33	50	87.11
1.21-1.30	33	90.00	37	91.22
1.31-1.40	39	94.33	31	94.67
1.41-1.50	13	95.78	15	96.33
1.51-1.60	11	97.00	8	97.22
1.61-1.70	8	97.89	6	97.89
1.71-1.80	7	98.67	9	98.89
1.81-1.90	4	99.11	4	99.33
1.91-2.00	7	99.89	2	99.56
2.01-2.10	1	100.00	2	99.78
2.11-2.20	0	100.00	2	100.00

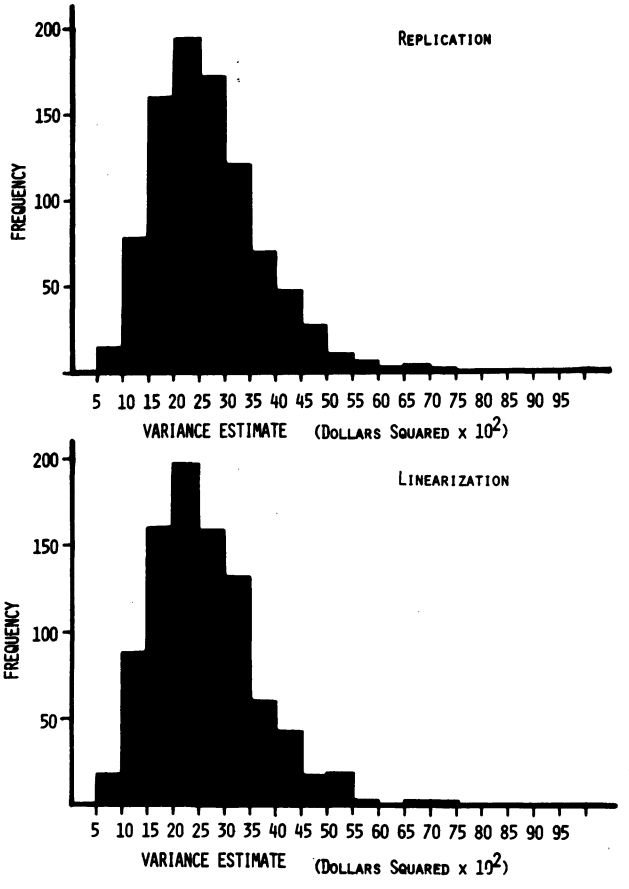

FIGURE 1: HISTOGRAMS OF 900 VARIANCE ESTIMATES AS CALCULATED BY THE BALANCED HALF-SAMPLE REPLICATION VARIANCE ESTIMATOR AND BY THE LINEARIZATION VARIANCE ESTIMATOR FOR THE POPULATION ESTIMATE AVERAGE INCOME PER PERSON FOR DESIGN III.

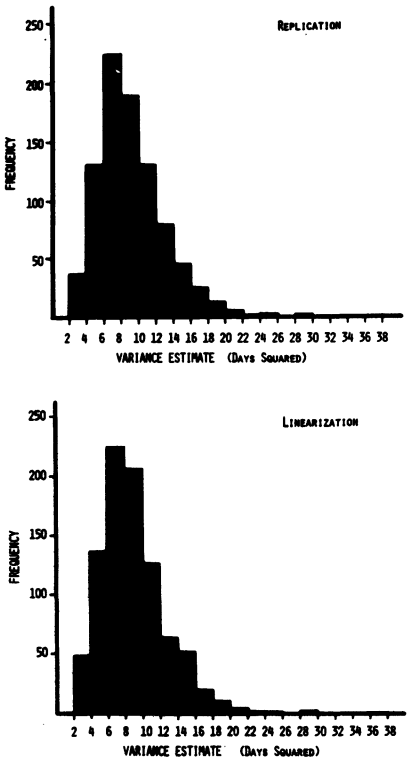

FIGURE 2: HISTOGRAMS OF 900 VARIANCE ESTIMATES AS CALCULATED BY THE BALANCED HALF-SAMPLE REPLICATION VARIANCE ESTIMATOR AND BY THE LINEARIZATION VARIANCE ESTIMATOR FOR THE POPULATION ESTIMATE THE AVERAGE NUMBER OF RESTRICTED ACTIVITY DAYS PER PERSON PER YEAR FOR DESIGN III.

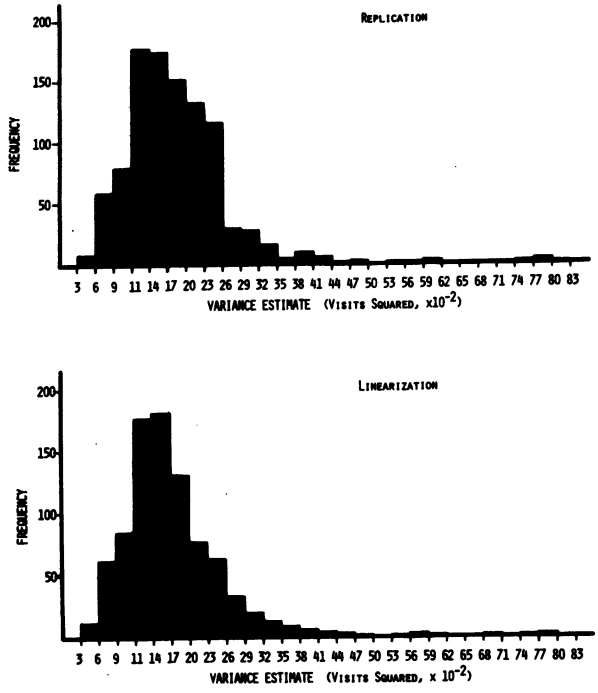

FIGURE 3: HISTOGRAMS OF 900 VARIANCE ESTIMATES AS CALCULATED BY THE BALANCED HALF-SAMPLE REPLICATION VARIANCE ESTIMATOR AND BY THE LINEARIZATION VARIANCE ESTIMATOR FOR THE POPULATION ESTIMATE AVERAGE NUMBER OF VISITS PER PERSON PER YEAR FOR DESIGN III.

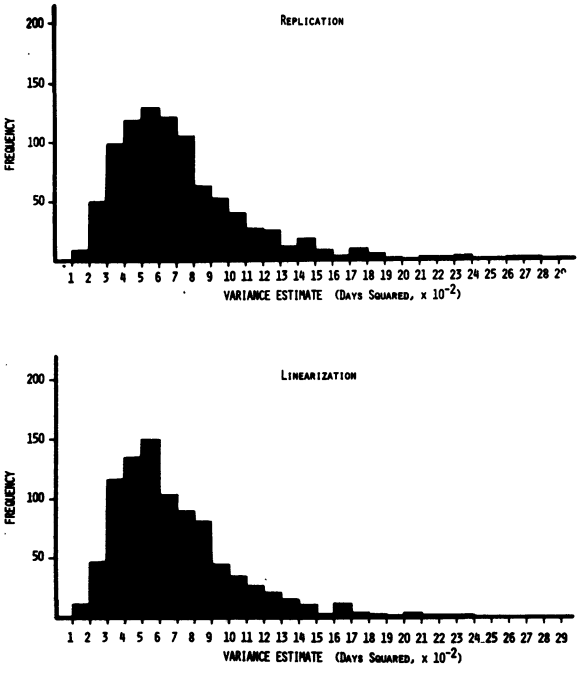

FIGURE 4: HISTOGRAMS OF 900 VARIANCE ESTIMATES AS CALCULATED BY THE BALANCED HALF-SAMPLE REPLICATION VARIANCE ESTIMATOR AND BY THE LINEARIZATION VARIANCE ESTIMATOR FOR THE POPULATION ESTIMATE AVERAGE NUMBER OF SHORT-STAY HOSPITAL DAYS PER PERSON PER YEAR FOR DESIGN III.

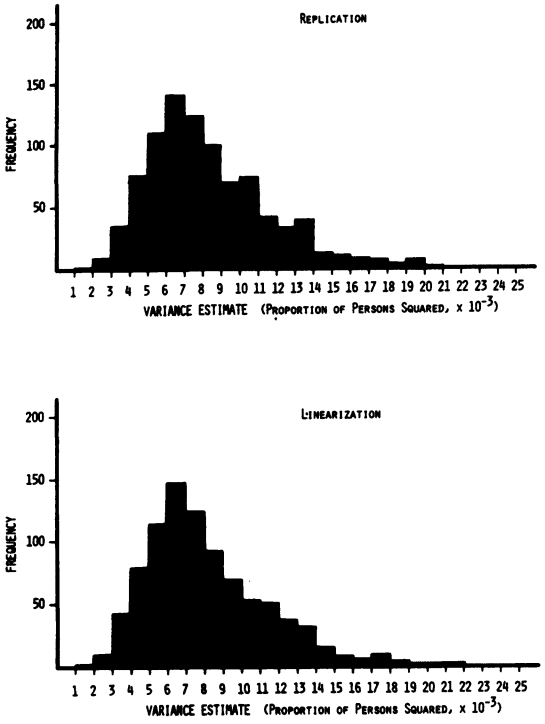

FIGURE 5: HISTOGRAMS OF 900 VARIANCE ESTIMATES AS CALCULATED BY THE BALANCED HALF-SAMPLE REPLICATION VARIANCE ESTIMATOR AND BY THE LINEARIZATION VARIANCE ESTIMATOR FOR THE POPULATION ESTIMATE PROPORTION OF POPULATION SEEING A PHYSICIAN IN LAST 12 MONTHS FOR DESIGN III.