ESTIMATING VARIANCE COMPONENTS FOR A COMPLEX SURVEY

Lynn Weidman and Todd Williams*
Bureau of the Census, Washington, DC 20233

1. Introduction

The Bureau of the Census conducts several major household surveys which measure various characteristics of the U.S. population. Each of these surveys has a similar design in which some housing units are selected in two stages and some in one stage from counties or groups of counties called primary sampling units (PSUs). This sampling is carried out so that each housing unit in the nation has the same overall probability of selection.

In particular, this research is concerned with the National Crime Victimization Survey (NCVS). The PSUs within each of four geographical regions are grouped into strata of roughly equal size (population). PSUs that are too large to fit into these strata are selected with probability one (self-representing or SR). In a one-PSU-per-stratum design like NCVS, the probability of a non-self-representing (NSR) PSU being in sample is the size of the PSU divided by its stratum size. Within each sample PSU housing units are selected without replacement in small geographically proximate groups called segments. Each person in a selected housing unit is interviewed seven times at six month intervals, with the first interview used only to bound the reference period for future interviews. Several stages of weighting, including adjustment for household nonresponse and control of specified age x race x sex combinations to projected national totals, are performed to convert the initial household probability of selection into final person weights for use in estimating national crime totals.

As a result of the procedures for selecting households and calculating final weights in NCVS and the other major household surveys, complex dependencies between households and between persons are introduced so that the usual textbook formulae for estimating variances cannot be directly applied. At various times at the Census Bureau, the alternative methods of linearization, random groups, jackknife and half sample replication have been investigated for possible use as a method for variance estimation in one or more of these household surveys.

A method that has been used for a variety of projects at the Bureau is half sample replication with a modification of the usual weights, as discussed in Dippo, Fay and Morganstein (1984). This modified half sample method was recently used to estimate variances of estimates of crime incidences for the NCVS. Additional research is planned to determine if for a given variable there is a relationship between its design effect and its between and within PSU variance components as a proportion of the total variance. A few of the possible uses of this information are: input to future design considerations, such as whether sample reduction should be accomplished through elimination of NSR PSUs or sample within PSUs; identification of ways in which variables should be grouped for estimating different generalized variance functions, rather than using a single overall generalized variance function; and supplying advice to public-use file users on approximating variances for their own modeling and data analytic studies.

One question that arose from this planned research was which method to use to get the "best" estimates of the variance components, since it is not necessarily the same one that is "best" for the total variances. If a relatively complicated procedure is judged to be "best", it may require extensive knowledge of the sample design and/or weighting, most of which is not available to users of the data outside the Census Bureau. In this case, is there a simpler procedure not requiring this knowledge, and generally available to all users of NCVS data, that gives acceptable variance estimates?

In this paper we will examine total variances and variance components computed by stratified jackknife (SJ), half sample (HS) and modified half sample (MHS) methods for the NCVS. The following section describes the variance estimation methods. Section 3 explains the comparisons performed and the general results. The tables include comparisons for a selection of estimates and are necessarily limited by the space available.

2. Variance Estimation Methods

We are estimating variances for crimes reported during interviews taking place in 1988, called collection year 1988. (It is simpler to work with collection year instead of calendar year data, and the variances will be similar.) The computational set-up of each of the methods will be briefly described, assuming that the reader is familiar with the basic jackknife and half sample procedures. In order to use common terminology in this description, the reader should think of the jackknife as a replication procedure.

First we define what we call the standard error computing units (SECUs) and pseudostrata for the methods. A SECU is a group of segments in which each record (person or household) has its weight multiplied by a common factor within each replicate of an estimation procedure. The SECUs and the factors they receive vary somewhat depending on whether we are calculating total, within SR PSU, within NSR PSU or between NSR PSU variances. Pseudostrata are collections of SECUs. Some of the details of SECU and pseudostrata construction follow and are summarized in Table 1.

For each SJ replicate one SECU is omitted, but only the weights of the remaining SECUs in the *same pseudostratum* (unlike the regular jackknife) are weighted by the factor

(number of SECUs in pseudostratum)/
(number of SECUs in pseudostratum - 1),

while the other segments retain their original weights (factors=1). The number of SJ replicates varies with the component being calculated, being equal to the number of

SR segments plus NSR PSUs, SR segments, NSR PSUs and NSR segments for the total, within SR, between NSR and within NSR components, respectively.

A common set of 164 completely balanced replicates is used in calculating each of the components by both HS and MHS. The SR segments are assigned to 80 pairs of SECUs and the 153 NSR PSUs to 75 pairs and one triplet, these groupings being the pseudostrata. (For simplicity of exposition we will treat the triplet as if it were a pair in the remainder of the paper.) A 164 x 164 Hadamard matrix is used to define the replicates.

In calculating total variance by HS, for each replicate one SECU in each pseudostratum gets the factor 2 and the other the factor 0. The corresponding factors for MHS are 1.5 and .5 for SR SECUs, and $F_{\alpha 1}$ and $F_{\alpha 2}$, defined next, for NSR SECUs.

For a particular NSR pseudostratum g containing two PSUs, define the following variables.

A_g = the total measure of size of pseudostratum g
A_{g1} = the total measure of size of PSU 1 in pseudostratum g
A_{g2} = the total measure of size of PSU 2 in pseudostratum g
$m_{\alpha g}$ = the Hadamard matrix entry for replicate α in pseudostratum g

Then PSU 1's replicate factor $F_{\alpha 1}$ for replicate α is

$$F_{\alpha 1} = \begin{cases} 2 - \frac{A_{g1}}{A_g} & \text{if } m_{\alpha g} = 1 \\ 1 - \frac{A_{g2}}{A_g} & \text{if } m_{\alpha g} = -1 \end{cases}$$

and PSU 2's replicate factor $F_{\alpha 2}$ for replicate α is

$$F_{\alpha 2} = \begin{cases} 2 - \frac{A_{g2}}{A_g} & \text{if } m_{\alpha g} = -1 \\ 1 - \frac{A_{g1}}{A_g} & \text{if } m_{\alpha g} = 1 \end{cases} .$$

When estimating total variances the MHS procedure is supposed to improve on the stability of variance estimates from the basic HS method by this modification of the variability of the weighting factors between replicates. In addition, MHS uses information from the structure of the design to obtain $F_{\alpha 1}$ and $F_{\alpha 2}$. One of the questions we are interested in here is whether there really are noticeable differences between HS and MHS for the computed variance estimates.

Another aspect of the HS and MHS methods we will examine is the effect of reweighting. The basic way of using these methods is to apply the factors to the final weights, which we call unreweighted and denote it by adding -U to a method's abbreviation. The strictly correct way to use these methods is to repeat the entire weighting procedure after applying the factors to the "initial" weights. We do this in a simplified form by repeating only the final stage of weighting, after multiplying each weight entering this stage by its appropriate factor. We call this reweighted and denote it by adding -R to a method's abbreviation. (Reweighting is not performed for SJ because of its large number of replicates and the consequent amount of storage space and computer time that would be required. Also note that for total variances only a fraction of the weights get a factor other than 1.0 in each replicate, while in the half-sample methods all weights get a factor other than 1.0.) Since only a small proportion of the population and any demographic group will experience any given type of crime, we expect that reweighting will have little effect. (See Ernst and Williams, 1987, for an examination of the effect of various factors, including different forms of $F_{\alpha 1}$ and $F_{\alpha 2}$ and reweighting, on variance estimates for the Current Population Survey.)

3. Comparisons

Variances were estimated for estimates of the total number of incidences in collection year 1988 for 31 types of crime (TOCs), 20 personal and 11 household. These are listed in Table 2. Note that most of them are constituents of others, as indicated by indentions in the listing. Variances were also estimated for five crime types broken down by specified demographic characteristics and geography, and for Hispanics within the demographic classifications. These breakdowns are listed in Table 3.

First we will look at the estimates of total variances (Table 4) for the TOCs, as well as for the specified estimates within the various demographic and geographic breakdowns.

3.1 Total Variances

Table 4 presents the estimates of total crime incidences and their estimated variances using MHS-R, and the ratios of the variance estimates obtained by the other methods to these variances. MHS-R is used as the basis of comparison because it makes use of the most design and weighting information. These comparisons can be summarized as follows.

TOC 1-20: HS-R and HS-U ratios are both larger and smaller than 1.0 with no particular pattern. SJ ratios are usually larger than 1.0. The three that are smaller are less than .868, which are quite extreme results. MHS-U ratios are usually slightly larger than 1.0. Note that the three largest ratios (>1.2) are for the three largest estimates and the three less than 1.0 are for the three smallest estimates.

TOC 501-511: HS-R and HS-U ratios are usually slightly below 1.0, with all of them being greater than .89. SJ ratios behave the same as for person crimes, where the ratios less than 1.0 are quite a bit less. MHS-U ratios are on both sides of 1.0, but four of them are greater than 1.13.

When the same comparisons are carried out for the estimates within the demographic and geographic groups, there is no overall pattern in the direction of the variability of the ratios about 1.0 for the other four methods. However, the ratios for SJ again tend to have a much wider range than the others.

There are two obvious conclusions to be made from these results. The first is that the stratified jackknife behaves differently than the half-sample methods, at least for the overall estimates. The cause for this behavior will have to be investigated further. (It may be partially due to SJ not being reweighted, but the effect of reweighting would have to be much larger than for the half sample methods to greatly reduce the variability of these ratios.) Secondly, on the whole, the simplest half-sample method, HS-U, gives variances similar to MHS-R. Thus if a data user is interested in estimating crime incidence variances using a half-sample approach, he is just about as well off using readily available half-sample software and not worrying about the design information needed for MHS and the weighting details needed for reweighting.

3.2 Variance Components

Next we look at how the total variance is partitioned into within SR, within NSR and between NSR PSU components. These component proportions are given in Table 5 for selected major crime types, which are representative of the patterns for all TOCs.

Before discussing the results we note how the computations were done. SR, total NSR and within NSR variances were computed directly, while the between NSR PSU component was obtained by subtraction. (If the within NSR component is greater than the total NSR component, then in Table 5 the between NSR proportion is given as 0.0.) SJ proportions sum to 1.0 because the total variance replicates are just the sum of the SR and total NSR replicates. For the unreweighted half sample methods, the sum of the proportions will differ from 1.0 because of the confounding in the total variance computations. E.g., since each reweighted SR PSU component variance was calculated using the total crime estimate for each replicate, not just the estimate for SR PSUs, it includes contributions from replicate factors in SR segments and from the effect the reweighting has on estimates of total NSR incidents.

Examination of Table 5 reveals that the behavior of the proportions is very similar for all the methods. (There are five lines for each TOC in Table 5, one for each of the estimation methods. In order they are MHS-R, HS-R, SJ, MHS-U, HS-U.) Only five of the SR proportions differ from those for MHS-R by more than 10%, and three of these are for TOC 1. Also 5 (TOCs 3,5,11,13,16) of the NSR-W proportions differ by more than 10%, but in this case they are all SJ. Results for household crimes and the demographic and geographic subgroups are very similar to those in Table 5. Again the obvious conclusion is that HS-U will give very similar results to the more complicated MHS-R, and SJ should be avoided because of occasional large differences.

4. Conclusions

This strictly computational comparison of variance estimation methods has shown that they give very similar results for NCVS and the crime incidence methods investigated. Users of these data in general would be just as well off using the relatively easier method of regular half samples without reweighting. In using these results it is important to remember that very small proportions of the total population and population subgroups have evidence of each of the types of crime. Further understanding of the results requires a more theoretical comparison of the methods. This should also show why SJ tends to show a somewhat different pattern of estimates for total variances, but not so much difference for the proportions.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

References

Bean, J.A and Schnack, G.A. (1977), "Components of Variance by Replicated BRR," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 200-203.

Casady, R.J. (1975), "The Estimation of Variance Components Using Balanced Repeated Replication," *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 352-357.

Dippo, C.S, Fay, R.E. and Morganstein, D.H. (1984), "Computing Variances from Complex Samples with Replicate Weights," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 489-494.

Ernst, L.R. and Williams, T.R. (1987), "Some Aspects of Estimating Variances by Half-Sample Replication Methods," Report CENSUS/SRD/RR-87/16 , U.S. Bureau of the Census, Statistical Research Division.

Folsom, R.E., Bayless, D.L. and Shah, B.V. (1971), "Jackknifing for Variance Components in Complex Sample Survey Design," *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 36-39.

Schindler, E. and Kulpinsky, S. (1981), "Components of Variance Replicated by BRR," *Proceedings of the Section on Survey Methods Research, American Statistical Association*, pp. 200-203.

Wolter, K.M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.

Table 1: Summary of Computational Details for the Variance Estimation Methods

Component	SR/NSR	Stratified Jackknife		Both Half Sample		
		SECU	Stratum	SECU	HS Factors	MHS Factors
Total	SR	Segment	PSU	Group of segments	0, 2	.5, 1.5
	NSR	PSU	PSU pair	PSU	0, 2	$F_{\alpha 1}$, $F_{\alpha 2}$
Within SR	SR	Segment	PSU	Group of segments	0, 2	.5, 1.5
	NSR	--[1]	--[1]	--[2]	1	1
Between NSR	SR	--[1]	--[1]	--[2]	1	1
	NSR	PSU	PSU pair	PSU	0, 2	$F_{\alpha 1}$, $F_{\alpha 2}$
Within NSR	SR	--[1]	--[1]	--[2]	1	1
	NSR	Segment	PSU	Half the segments in a PSU	0, 2	.5, 1.5

--[1] No replicates are formed by omitting segments from these PSUs. Factors are 1 for all segments in all replicates.
--[2] SECUs need not be defined. Factors are 1 for all segments in all replicates.

Table 2: Types of Crime

1. Crimes against persons
2. Crimes of violence
3. Rapes
4. Robberies
5. Completed robberies
6. Completed robberies with injury
7. Completed robberies without injury
8. Attempted robberies
9. Attempted robberies without injury
10. Assaults
11. Aggravated assaults
12. Completed (with injury) aggravated assaults
13. Simple assaults
14. Attempted simple (without weapon) assaults
15. Crimes of theft
16. Personal larcenies with contact
17. Purse snatchings
18. Pocket pickings
19. Personal larcenies without contact
20. Completed personal larcenies without contact, value <&50
501. Crimes against households - property
502. Burglaries
503. Burglaries, forcible entry
504. Burglaries, unlawful entry
505. Burglaries, attempted forcible entry
506. Household larcenies
507. Completed household larcenies
508. Completed household larcenies, value < $50
509. Completed household larcenies, value $50 and over
510. Motor vehicle thefts
511. Completed motor vehicle thefts

Table 3: Demographic and Geographic Breakdowns for Crimes

A. Assaults, personal larcenies without contact for persons who are nonblack, black, female, age 12-24, age 65+, never married, in family with income < $10,000
B. Burglaries, completed household larcenies, motor vehicle thefts where head of household is nonblack, black, female, age 12-34, age 65+, never married, in family with income < $10,000
C. A and B for in metropolitan area, in central city
D. Table 2, A and B for Hispanics

Table 4: Comparison of Total Variances for Table 2 Crimes

TOC	Estimate MHS-R	Variance MHS-R	Proportion of MHS-R Variance HS-R	SJ	MHS-U	HS-U
1	0.1957D+08	0.1020D+12	0.972	1.176	1.210	1.084
2	0.5919D+07	0.2713D+11	0.970	0.957	1.059	0.991
3	0.1487D+06	0.3487D+09	1.023	0.986	0.979	0.986
4	0.1010D+07	0.2738D+10	1.015	0.908	0.983	0.989
5	0.6407D+06	0.1643D+10	1.007	0.866	0.981	0.976
6	0.2494D+06	0.4860D+09	1.000	1.093	0.994	0.981
7	0.3912D+06	0.1087D+10	1.021	0.791	0.971	0.972
8	0.3694D+06	0.8759D+09	1.021	1.004	0.994	0.997
9	0.2723D+06	0.6718D+09	1.026	0.958	0.995	1.014
10	0.4760D+07	0.2161D+11	0.957	0.973	1.074	0.990
11	0.1732D+07	0.6219D+10	0.952	0.948	1.047	0.982
12	0.5909D+06	0.1458D+10	0.990	0.891	1.029	1.010
13	0.3027D+07	0.1015D+11	0.971	1.074	1.077	1.013
14	0.2135D+07	0.6498D+10	0.981	1.087	1.068	1.021
15	0.1365D+08	0.5491D+11	0.992	1.192	1.189	1.085
16	0.4745D+06	0.9591D+09	0.998	1.066	1.010	0.979
17	0.1494D+06	0.2432D+09	0.980	1.160	0.992	0.957
18	0.3250D+06	0.6655D+09	1.010	1.073	1.002	0.984
19	0.1318D+08	0.5309D+11	0.992	1.193	1.186	1.090
20	0.5120D+07	0.1811D+11	1.001	1.127	1.051	1.005
501	0.1601D+08	0.6417D+11	0.985	1.041	1.114	1.004
502	0.5908D+07	0.1989D+11	0.995	0.936	0.994	0.935
503	0.2029D+07	0.6048D+10	1.010	0.881	0.930	0.905
504	0.2614D+07	0.6998D+10	1.005	1.075	1.017	0.989
505	0.1265D+07	0.3016D+10	0.990	0.994	1.022	0.979
506	0.8507D+07	0.2985D+11	0.988	1.037	1.109	1.028
507	0.7929D+07	0.2704D+11	0.986	1.025	1.112	1.027
508	0.3282D+07	0.9589D+10	0.996	0.955	1.051	0.997
509	0.4124D+07	0.1180D+11	0.975	1.009	1.094	1.029
510	0.1597D+07	0.5135D+10	1.004	0.751	1.015	1.001
511	0.1053D+07	0.2646D+10	1.008	0.869	1.019	1.011

Table 5: Comparison of Variance Components for Selected Personal Crimes[1]

TOC	Estimate MHS-R	Variance	Proportion of Total SR	NSR-W	NSR-B
1	0.1957D+08	0.1020D+12	0.264	0.217	0.518
		0.9920D+11	0.275	0.225	0.496
		0.1199D+12	0.367	0.207	0.425
		0.1234D+12	0.296	0.213	0.490
		0.1106D+12	0.331	0.238	0.431
3	0.1487D+06	0.3487D+09	0.736	0.182	0.074
		0.3566D+09	0.721	0.178	0.077
		0.3438D+09	0.763	0.219	0.018
		0.3414D+09	0.768	0.182	0.050
		0.3437D+09	0.763	0.181	0.056
5	0.6407D+06	0.1643D+10	0.836	0.141	0.020
		0.1655D+10	0.834	0.141	0.014
		0.1424D+10	0.825	0.175	0.000
		0.1612D+10	0.842	0.151	0.007
		0.1604D+10	0.847	0.152	0.001
8	0.3694D+06	0.8759D+09	0.761	0.227	0.000
		0.8945D+09	0.750	0.219	0.000
		0.8791D+09	0.766	0.214	0.020
		0.8702D+09	0.782	0.218	0.000
		0.8732D+09	0.779	0.221	0.000
11	0.1732D+07	0.6219D+10	0.473	0.431	0.101
		0.5918D+10	0.500	0.456	0.050
		0.5895D+10	0.489	0.351	0.160
		0.6514D+10	0.490	0.430	0.080
		0.6107D+10	0.523	0.458	0.019
13	0.3027D+07	0.1015D+11	0.466	0.367	0.173
		0.9857D+10	0.483	0.380	0.144
		0.1090D+11	0.495	0.322	0.182
		0.1093D+11	0.457	0.348	0.194
		0.1028D+11	0.486	0.370	0.143
16	0.4745D+06	0.9591D+09	0.767	0.226	0.000
		0.9576D+09	0.771	0.210	0.000
		0.1022D+10	0.802	0.198	0.000
		0.9691D+09	0.760	0.240	0.000
		0.9387D+09	0.785	0.215	0.000
19	0.1318D+08	0.5309D+11	0.327	0.236	0.433
		0.5266D+11	0.332	0.240	0.418
		0.6335D+11	0.400	0.213	0.387
		0.6297D+11	0.324	0.229	0.448
		0.5787D+11	0.352	0.249	0.399

[1] In the headings W denotes within PSU and B denotes between PSU. There are 5 rows for each TOC, one for each of the estimation procedures. They are in the order MHS-R, HS-R, SJ, MHS-U, and HS-U.