

1997 NSAF Variance Estimation

Report No. 4

Prepared by:

Ismael Flores-Cervantes

J. Michael Brick

Ralph DiGaetano.



Assessing
the New
Federalism

*An Urban Institute
Program to Assess
Changing Social Policies*

NSAF Methodology

Preface

1997 NSAF Variance Estimation is the fourth report in a series describing the methodology of the 1997 National Survey of America's Families (NSAF). It has been recently reissued as a companion to the 1999 NSAF Report on the same subject (also No. 4 in that series). Two closely related reports are those for 1997 and 1999 on the design and estimation carried out in both rounds. These appear as Report No. 2 and 4 in each series. For the 1997 survey, see also Report No. 14.

About the National Survey of America's Families (NSAF)

As discussed elsewhere (e.g., see especially Report No. 1 in the 1997 NSAF methodology series), NSAF is part of the Assessing the New Federalism Project at the Urban Institute, being done in partnership with Child Trends. Data collection for the NSAF was conducted by Westat.

In each rounds of NSAF, carried out so far, over 40,000 households were interviewed, yielding information on over 100,000 people. NSAF has focused on the economic, health, and social characteristics of children, adults under the age of 65, and their families. The sample is representative of the nation as a whole and of 13 states shown below. Because of its large state sample sizes, NSAF has an unprecedented ability to measure differences between the 13 states it targeted.

About the 1997 and 1999 NSAF Methodology Series

The 1997 and 1999 methodology series of reports have been developed to provide readers with a detailed description of the methods employed to conduct the 1997 NSAF. The two series are nearly parallel, except for the documentation of the public use files, where an on-line system is being used for the 1999 survey and we are planning to reissue the 1997 files on a similar basis.

Report No 1 in the 1997 series introduces NSAF. Report Nos.2 through 4 in both series—plus Report No. 14 in the 1997 series—describe the sample design, how survey results were estimated and how variances were calculated. Report Nos. 5 and 9 in each series describe the interviewing done in for the telephone (RDD) and in-person samples. Report Nos. 6 and 15 in the 1997 series and Report No. 6 in the 1999 series displays and discusses the comparisons we made to surveys that partially overlapped NSAF in content—including the Current Population Survey and the National Health Interview Survey, among others. Report Nos. 7 and 8 in both series cover what we know about nonresponse rates and nonresponse biases. Report No. 10 in both series covers the details of the survey processing, after the fieldwork was completed, including the imputation done for items that were missing. Report No. 11 in both series introduces the public use files made available.

In the 1997 series, there were additional reports on the public use files available in a PDF format as Report No. 13, 17-22. These will all eventually be superceded by the on-line data file codebook system that we are going to employ for the 1999 survey. The 1997 and 1999 NSAF questionnaires are available respectively as Report No. 12 in the 1997 series and Report No. 1 in the 1999 series. Report No. 16 for the 1997 series, the only report not so far mentioned contains occasional papers

of methodological interest given at professional meetings through 1999, regarding the NSAF work as it has progressed over the years since 1996 when the project began.

About this 1997 Report

Report No. 4 describes the methods and results of computing sampling errors for the 1997 National Survey of America's Families (NSAF). First, an overview of the sample design and summaries of the precision of the survey estimates for both estimates of children and adults are presented. The second chapter presents a general review of the two main methods of computing sampling errors or variances of estimates from surveys with complex survey designs like the NSAF. The third chapter discusses why the replication method of variance estimation was chosen as the main method for the NSAF and describes procedures for computing replicate estimates of variance from the data. The fourth chapter describes how software available for computing sampling errors can be used with the data. The final chapter summarizes the findings and methods used.

For More Information

For more information about the National Survey of America's Families, contact Assessing the New Federalism, Urban Institute, 2100 M Street, NW, Washington, DC 20037, telephone: (202) 261-5377, fax: (202) 293-1918, Website: <http://newfederalism.urban.org>. For information about this report, contact BrickM1@Westat.com.

Jenny Kenney
and
Fritz Scheuren

TABLE OF CONTENTS

<u>Chapter</u>		<u>Page</u>
1	Introduction	1-1
	1.1 Overview of the Survey	1-1
	1.2 Design Effects	1-2
	1.3 Structure of the Report	1-10
2	Methods for Variance Estimation	2-1
	2.1 Replication Methods	2-1
	2.2 Taylor Series Methods	2-2
3	Replicaton for the NSAF	3-1
	3.1 Choice of Method	3-1
	3.2 Design of Replicates	3-2
	3.3 Issues Related to Replicate Structure	3-3
4	Software for Computing Variances	4-1
	4.1 WesVar	4-1
	4.2 SUDAAN	4-1
	4.3 STATA	4-3
	4.4 Comparison WesVar, SUDAAN and STATA	4-4
5	Summary	5-1
	References	R-1

List of Appendices

<u>Appendix</u>		<u>Page</u>
A	Wesvar Example	A-1
B	Sudaan Example	B-1
C	Stata Example	C-1

TABLE OF CONTENTS (CONTINUED)

List of Tables

<u>Table</u>		<u>Page</u>
1-1	Average DEFF and DEFT for Estimates from the Adult Pair File for All Adults and Low-income Adults, by Site.....	1-6
1-2	Average DEFF and DEFT for Estimates from the Adult Pair File for All “Option B” and Low-income “Option B” Adults, by Site.....	1-6
1-3	Average DEFF and DEFT for Estimates from the Adult Pair File for All Hispanics and Low-income Hispanics, by Site.....	1-7
1-4	Average DEFF and DEFT for Estimates from the Adult Pair File for All Blacks and Low-income Blacks, by Site	1-7
1-5	Average DEFF and DEFT for Estimates from the Adult Pair File for All Adults in Households with Children and for Low-income Adults in Households with Children, by Site.....	1-8
1-6	Average DEFF and DEFT for Estimates from the Adult Pair File for All adults in Households with No Children and for Low-income Adults in Households with No Children, by Site	1-8
1-7	Average DEFF and DEFT for Estimates from the Child File for All Children and Low-income Children, by Site	1-9
1-8	Average DEFF and DEFT for Estimates from the Child File for All Hispanic Children and Low-income Hispanic Children, by Site.....	1-9
1-9	Average DEFF and DEFT for Estimates from the Child File for All Black Children and Low-income Black Children, by Site	1-10
4-1	Estimates of Children Computed Using WesVar, SUDAAN, and STATA.....	4-6

List of Figures

<u>Figure</u>		<u>Page</u>
1-1	Study Areas	1-1
3-1	Ratio of DEFFs with Initial Variance Units to DEFFs with Revised Variance Units.....	3-5

Chapter 1

Introduction

This report describes the methods and results of computing sampling errors for the 1997 National Survey of America's Families (NSAF). The first chapter gives an overview of the sample design and summaries of the precision of the survey estimates for both children and adults.

The remainder of the report describes the methodology used to create these types of estimates of sampling variability. Chapter 2 is a general review of the two main methods of computing sampling errors or variances of estimates from surveys with complex survey designs like the NSAF. The third chapter discusses why the replication method of variance estimation was chosen as the main method for the NSAF and describes procedures for computing replicate estimates of variance from the data. Chapter 4 and its attachments show how software available for computing sampling errors can be used with the data. The final chapter summarizes the findings and methods used.

1.1 Overview of Survey

The NSAF collected information on the economic, health, and social dimensions of the well-being of children, adults under the age of 65, and their families in 13 states, Milwaukee, and the balance of the nation (see figure 1-1). In this section we briefly outline the sample design features. The details on the design are given in the *1997 NSAF Sample Design Report*, another report in this series.

Figure 1-1
Study Areas

Alabama	Massachusetts	New Jersey	Milwaukee County
California	Michigan	New York	Balance of Wisconsin
Colorado	Minnesota	Texas	Balance of Nation
Florida	Mississippi	Washington	

The primary goal of the survey was to obtain social and economic information about children in low-income households (income below 200 percent of the poverty threshold). Similar data on children in all households, low-income adults under age 65, and on other adults under age 65 were also obtained.

The two components of the survey were a random digit dialing (RDD) survey of households with telephones, and an area sample conducted in person for those households without telephones. This dual-frame approach is further described in Waksberg, et al. (1997).

In the RDD sample, screener-based subsampling of households was used to sample low-income households at a higher rate than other households. A very short income question was asked during the RDD screening interview, and those that reported an absence of children or reported incomes above 200 percent of the poverty threshold were subsampled. In the area sample,

blocks with very high telephone coverage rates as of the 1990 Census were excluded to reduce costs.

Within both the RDD and the area samples, household members were subsampled to reduce the respondent burden. If there were multiple children under age 6, one was randomly selected. The same was done for children 6 to 17 years old. Data were collected from the most knowledgeable adult (MKA) in the household for the sampled child. During the MKA interview, data were also collected for the MKA and his/her spouse/partner. Most questions asked about the MKA were repeated in reference to the spouse/partner; however, some questions on health insurance and health care utilization were asked in reference to only one of the two. The target of these questions was randomly assigned to either the MKA or his/her spouse/partner. Some questions relating to feelings, religious activities, and opinions were asked only about the MKA.

Other adults in households with children were subsampled, as were adults in adult-only households. Adults were eligible only if they would not have been MKAs for other children in the household if those children had been selected. Self-response was required for sampled adults, and data were also collected about his/her spouse/partner (if living in the same household). Data were not collected directly from the spouse of a sampled adult. As in the MKA interview, there were also some questions relating to feelings, religious activities, and opinions that were asked only about the sampled adult.

1.2 Design Effects

In order to evaluate the precision of the sample estimates derived from a survey, sampling errors are computed from the sample. Estimates of sampling errors can be used to make inferences about the size of the difference between two population parameters based on the values of corresponding sample estimates, their estimated precision, and the expected probability distribution of such a difference. For instance, one could compare the proportions of persons in poverty in two regions of the country by computing the difference of the estimated proportions while taking into account the estimated sampling error of this difference. Estimated sampling errors can appear in several forms, including the estimated standard error of a single survey estimate, the ratio of the estimated standard error to the single survey estimate called the coefficient of variation of the survey estimate, and a range of values, or confidence interval, about the survey estimate.

Another way of describing the variability of an estimate from a survey is by using the “design effect.” The term design effect is used to describe the variance of sample estimates for a particular sample design relative to the corresponding variance of a simple random sample with the same sample size. Design effects are used to evaluate the efficiency of the sampling design and estimation procedure utilized to develop the estimates.

The concept of design effect was popularized by Kish (1985) to deal with complex sample designs involving stratification and clustering. Stratification generally leads to a gain in efficiency over simple random sampling, but clustering usually leads to deterioration in the efficiency of the estimate due to positive intracluster correlation among the subunits in the clusters. In order to determine the total effect of any complex design on the sampling variance in

comparison to the alternative simple random sampling, one calculates a ratio of variances associated with an estimate, namely

$$DEFF = \frac{\text{sampling variance of a complex sample}}{\text{sampling variance of a simple random sample}}$$

This ratio is called the design effect (DEFF) of the sampling design for the estimate. This ratio measures the overall efficiency of the sampling design and the estimation procedure utilized to develop the estimate. At the analysis stage, the DEFF is useful because most statistical analysis software (such as SAS and SPSS) assume the data are from a simple random sample when computing sampling errors of estimates. The DEFF can, in some circumstances, indicate how appropriate this is, and can be used to adjust these simple estimates to produce ones that are closer to the actual sampling errors of the estimates (Skinner, Holt and, Smith 1989).

The design effect for a proportion is estimated as

$$DEFF_{PROP} = \frac{v(p)_{COMPLEX}}{v(p)_{SRS}}$$

where p denotes the estimate of the proportion,

$v(p)_{SRS}$ is the estimated simple random variance $v(p)_{SRS} = \frac{p(1-p)}{n}$, and
 $v(p)_{COMPLEX}$ is the variance of the complex sample calculated appropriately.

In most cases, design effects for complex samples are larger than one. In the NSAF, design effects are greater than one because of differential sampling fractions and the intracluster correlation of units within clusters (Kish 1992). As will be seen shortly, some design effects from the survey are considerably greater than one.

In the NSAF and most large-scale surveys, a large number of data items or variables are collected from respondents. Each variable has its own design effect. One way to represent all of these is to compute design effects for a number of similar variables and then average them. This average is used to represent the design effects for the group of variables. This practice is described in Wolter (1985).

Tables 1-1 to 1-6 show average design effects for 24 estimates of proportions for adults from the adult pair file. The tables correspond to all adults, "Option B" adults (those who are not MKAs or the spouse of an MKA), Hispanics, blacks, adults in households with children, and adults in households without children. Each of the tables also has two sets of estimates, one for all adults and the other for low-income adults (less than or equal to 200 percent of the poverty level). Tables 1-7 to 1-9 present corresponding design effects for 33 estimates of children. The child tables are for all children, Hispanics and blacks. Each table has estimates for all children and for low-income children.

Each of the tables has four entries for each site and for all and low-income persons. The first column is the average DEFF, the second is the maximum DEFF, the third is the minimum DEFF, and the fourth is called the DEFT. The DEFT is the square root of the design effect, so it is like the DEFF but on the scale of the standard deviation of the estimate rather than the variance. The figures labeled DEFT in the tables are actually the average of the DEFTs.

Verma, Scott, and O’Muirheartaigh (1980) discuss the use of the DEFT. It is a convenient measure because it can be used directly in computing confidence intervals for the estimates, whereas the square root of the DEFF must be computed before it can be used. However, the main reason for presenting the DEFTs in this application is because it dampens some of noise associated with the DEFFs. The maximum and minimum values of the DEFFs in the tables show that there is considerable variability in these quantities. By taking the square root of the DEFF and averaging these values, the variability is somewhat reduced. For example, in table 1-7, the average DEFF for Minnesota is 2.65, while the maximum is 14.14 and the minimum is 1.19. This value is unusually large given the other values in the table. The average DEFT of 1.56 is also large, but not as different from the values for the other sites.

Before discussing the tables in more detail, the most important factors that result in design effects larger than one in the NSAF are reviewed. These factors are:

1. Oversampling by study area. The need for both study area and national estimates required oversampling to produce stable separate estimates for the 13 specified sites. This oversampling increased the design effect for national estimates.
2. Household screening. Additional variability comes from the subsampling of households without children and those above 200 percent of poverty. The misclassification of households as “above 200 percent poverty” when they actually were not also increases the variance of estimates restricted to persons at or below 200 percent of the poverty level. See Flores-Cervantes, et al. (1998) for a discussion of this topic.
3. Within-household subsampling. Differential sampling rates at the person level also contribute to increases in design effects. Children and adults were subsampled within households for both the RDD and area sample components.¹
4. Clustering of households in the area sample. For the area sample component of the survey, households were clustered within segments and segments were clustered within PSUs. Design effects increase to the extent that respondents in the same cluster are similar in their responses to survey items. For estimates of low-income respondents, increased design effects due to the area sample clustering were more important because a larger percentage of low-income persons were in nontelephone households.

¹ Households without children and households above 200 percent of poverty were not subsampled in the area component of the study, so design effects associated with such subsampling were not incurred in the area sample but were for the RDD sample. However, the lower sampling rates used in the area sample increased sample variability for estimates where the RDD and area sample are pooled.

Table 1-1 shows that the average DEFT for estimates of all adults in each site was about 1.3 to 1.4, so the standard error of the estimate was about 30 to 40 percent greater than expected from a simple random sample. The average DEFT for the national estimate is significantly larger, 2.16, as expected because of the oversampling by site. The average DEFTs for the low-income adults in the table are slightly larger than the average DEFTs for all adults. The patterns by site and for the nation are similar to the ones for all adults. The main reason the low-income DEFTs are larger is because the misclassification of households by poverty status increased the variability in the weights for this subgroup.

Table 1-2 gives the same statistics but is restricted to “Option B” adults. The patterns are the same as in table 1-1 but the average DEFTs are smaller. The smaller DEFTs are due to two main factors: (1) nearly all “Option B” adults were selected at about the same rate while all adults include MKAs and “Option B” adults who were selected at different rates, and (2) the “Option B” adults have smaller cluster sizes than all adults and this reduces the variance of the estimate.

Table 1-3 and 1-4 give the estimates for adult Hispanics and blacks, respectively. The patterns here are very similar to those noted above, but these two subgroups have much smaller sample sizes. As a result some of the statistics used in preparing the averages may not be very stable. For example, 12 to 15 of the estimates used in computing the averages in these tables were based on sample sizes of less than 20 persons. This effect also shows up in the estimates of the maximum and minimum DEFFs. For example, in Florida the maximum DEFF for Hispanics is 8.05 and for low-income Hispanics it is 11.61. These are much higher than for any other statistics in the table. We discuss another reason for these large DEFFs below.

The last two tables of adults are for those who live in households with and without children. Once again, some very large values for the maximum DEFF are occurring. For example, the maximum DEFF for all adults in households with children for Minnesota is 12.22. All of the statistics with very large DEFFs that were used in the tables were examined. We found that living in public housing was one statistic that often had a large DEFF. This statistic is highly related to living in a nontelephone household and is thus subject to a large clustering effect. While these DEFF estimates are correct, including them in the average DEFTs tends to increase the averages. Because such estimates are very relevant to analyses of the data, the estimates were included in the averages rather than dropped.

Table 1-7 gives the average DEFTs for all children and for low-income children by site. The findings are again similar to those from the adult file. The average DEFT for all children across the sites is about 1.25 to 1.56, with the national DEFT being 2.29. For children in low-income families, the average DEFT is slightly larger, mirroring the result shown in table 2-1.

Table 1-8 and 1-9 give the average DEFTs for Hispanic and black children. The smaller sample size issue identified with adult Hispanics and blacks is even more serious for estimates of children. For example, 64 of the estimates used in computing the averages for the low-income Hispanics were based on sample sizes of less than 20 persons. The clustering for estimates such as the percent of children living in public housing also affects the DEFTs in these tables.

Table 1-1
Average DEFF and DEFT for Estimates from the Adult Pair File for All Adults
and Low-income Adults, by Site

Study Area	All Adults				Low-income Adults			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	2.09	3.25	1.37	1.44	2.24	4.11	0.92	1.47
Balance of Wisconsin	1.57	2.25	1.06	1.25	1.94	3.47	0.84	1.37
California	1.74	2.87	0.65	1.31	2.25	4.28	1.22	1.49
Colorado	1.78	2.29	1.27	1.33	1.95	3.09	1.19	1.39
Florida	2.06	3.35	0.83	1.42	2.62	4.95	1.84	1.60
Massachusetts	1.92	2.85	0.86	1.37	2.38	4.07	1.34	1.53
Michigan	1.77	3.61	0.92	1.31	2.14	3.51	0.83	1.45
Milwaukee	1.68	2.45	0.92	1.29	2.13	3.76	1.16	1.44
Minnesota	2.22	7.05	1.16	1.46	2.61	8.03	1.42	1.58
Mississippi	1.92	3.28	1.34	1.38	2.02	3.75	1.28	1.41
New Jersey	1.81	2.53	1.02	1.34	2.20	3.43	0.97	1.47
New York	1.93	2.82	0.59	1.37	2.45	3.53	1.07	1.55
Texas	2.10	3.26	0.56	1.43	2.58	4.60	1.25	1.58
Balance of the U.S.	1.71	2.80	0.90	1.29	1.99	3.23	1.26	1.39
Washington	1.73	3.07	1.03	1.31	2.01	3.08	1.18	1.40
National	4.74	8.20	1.71	2.16	5.21	8.90	3.09	2.26

Table 1-2
Average DEFF and DEFT for Estimates from the Adult Pair File for All
“Option B” and Low-income “Option B” Adults, by Site

Study Area	All				Low-income			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.54	2.24	0.99	1.24	1.62	2.81	0.94	1.26
Balance of Wisconsin	1.37	2.25	0.86	1.16	1.55	2.50	0.53	1.22
California	1.42	2.17	0.66	1.18	1.65	2.42	0.97	1.28
Colorado	1.46	2.32	0.95	1.20	1.54	2.37	0.88	1.23
Florida	1.34	2.09	0.74	1.15	1.60	3.06	1.16	1.26
Massachusetts	1.57	2.15	0.63	1.24	1.73	2.90	0.98	1.30
Michigan	1.49	2.56	0.85	1.21	1.72	2.68	0.69	1.29
Milwaukee	1.31	2.21	0.58	1.14	1.59	2.90	0.61	1.24
Minnesota	1.64	3.37	1.00	1.27	1.90	4.26	1.10	1.36
Mississippi	1.53	2.45	0.91	1.23	1.53	2.20	0.91	1.23
New Jersey	1.57	2.19	0.79	1.24	1.80	2.96	0.88	1.33
New York	1.43	2.52	0.71	1.18	1.69	2.45	0.58	1.29
Texas	1.38	2.25	0.43	1.16	1.60	3.14	0.74	1.25
Balance of the U.S.	1.42	2.29	0.85	1.18	1.50	2.05	0.97	1.22
Washington	1.34	2.35	0.98	1.15	1.47	2.46	0.81	1.20
National	3.66	5.09	1.97	1.90	3.77	5.81	2.53	1.93

Table 1-3
Average DEFF and DEFT for Estimates from the Adult Pair File for
All Hispanics and Low-income Hispanics, by Site

Study Area	All				Low-income			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.73	4.50	0.46	1.26	1.41	4.57	0.43	1.15
Balance of Wisconsin	1.99	3.61	0.35	1.37	2.23	4.88	0.42	1.39
California	2.09	3.45	0.91	1.43	2.32	5.03	1.43	1.50
Colorado	1.58	2.56	0.89	1.24	1.82	3.01	1.01	1.33
Florida	2.69	8.05	1.07	1.58	2.83	11.61	0.54	1.61
Massachusetts	2.24	3.73	1.27	1.48	2.14	3.32	0.74	1.44
Michigan	1.62	3.37	0.43	1.24	1.55	2.92	0.30	1.21
Milwaukee	1.90	2.73	0.40	1.35	1.94	4.24	0.45	1.35
Minnesota	1.64	2.99	0.56	1.25	1.99	3.68	0.57	1.37
Mississippi	1.30	3.52	0.68	1.11	1.21	3.76	0.20	1.05
New Jersey	1.96	2.94	0.86	1.38	2.05	3.32	0.84	1.41
New York	2.31	4.01	0.95	1.50	2.48	4.09	0.83	1.55
Texas	2.25	3.78	0.94	1.47	2.35	5.31	1.05	1.49
Balance of the U.S.	1.63	2.95	0.36	1.26	1.61	3.26	0.47	1.25
Washington	1.75	3.18	0.79	1.30	1.79	3.30	0.75	1.31
National	3.63	6.40	2.15	1.89	3.86	5.73	2.16	1.94

Table 1-4
Average DEFF and DEFT for Estimates from the Adult Pair File for
All Blacks and Low-income Blacks, by Site

Study Area	All				Low-income			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	2.06	4.29	0.96	1.41	1.93	2.91	0.79	1.37
Balance of Wisconsin	1.49	3.24	0.55	1.20	1.28	4.36	0.61	1.10
California	1.83	3.38	1.05	1.34	1.84	3.47	1.03	1.34
Colorado	1.25	2.29	0.25	1.09	1.26	1.74	0.43	1.11
Florida	2.46	4.70	0.34	1.53	2.44	4.74	0.47	1.53
Massachusetts	1.91	2.97	1.27	1.37	1.95	3.85	0.68	1.38
Michigan	1.73	3.51	0.86	1.30	2.01	3.57	1.15	1.40
Milwaukee	1.73	3.51	0.74	1.30	1.57	2.67	0.95	1.24
Minnesota	2.22	8.16	0.42	1.42	2.30	8.78	0.56	1.43
Mississippi	2.01	3.32	1.26	1.40	1.84	3.68	0.96	1.34
New Jersey	1.71	2.82	0.51	1.29	1.64	2.38	0.69	1.27
New York	1.80	2.70	0.75	1.33	1.79	3.32	0.76	1.32
Texas	2.18	4.64	0.23	1.44	2.13	3.99	0.46	1.43
Balance of the U.S.	1.79	3.63	0.94	1.32	1.66	2.73	1.03	1.28
Washington	1.39	2.02	0.83	1.17	1.37	1.98	0.92	1.16
National	5.16	9.30	2.73	2.25	4.90	8.11	3.17	2.20

Table 1-5
Average DEFF and DEFT for Estimates from the Adult Pair File for All
Adults in Households with Children and for Low-income Adults in Households
with Children, by Site

Study Area	All				Low-income			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.69	2.72	0.96	1.29	1.83	3.34	0.59	1.33
Balance of Wisconsin	1.53	2.47	0.85	1.22	1.81	3.39	0.98	1.32
California	1.95	4.13	0.97	1.38	2.32	4.74	0.94	1.49
Colorado	1.59	2.65	1.02	1.25	1.78	2.52	1.04	1.33
Florida	2.03	3.70	1.01	1.41	2.49	4.08	0.97	1.56
Massachusetts	1.74	2.39	0.77	1.31	2.17	3.35	1.47	1.47
Michigan	1.75	2.96	1.09	1.32	1.84	3.05	0.93	1.34
Milwaukee	1.81	3.41	1.13	1.34	2.37	5.34	1.31	1.51
Minnesota	2.47	12.22	0.80	1.49	2.79	12.59	0.64	1.59
Mississippi	1.96	3.57	1.20	1.39	1.95	3.71	1.09	1.38
New Jersey	1.70	2.70	0.91	1.29	1.94	2.93	1.27	1.38
New York	2.08	4.01	0.92	1.42	2.19	4.44	1.20	1.46
Texas	2.07	3.16	1.20	1.43	2.30	4.05	1.20	1.50
Balance of the U.S.	1.85	4.22	1.15	1.33	2.10	4.37	1.24	1.42
Washington	1.69	2.67	0.75	1.28	2.04	3.68	1.01	1.41
National	5.19	10.25	2.54	2.25	5.51	10.26	3.08	2.31

Table 1-6
Average DEFF and DEFT for Estimates from the Adult Pair File for All
Adults in Households with no Children and for Low-income Adults in Households
with No Children, by Site

Study Area	All				Low-income			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.56	2.30	1.10	1.24	1.59	2.79	0.64	1.25
Balance of Wisconsin	1.43	2.58	0.86	1.18	1.63	2.86	0.51	1.25
California	1.44	2.05	0.84	1.20	1.56	2.33	0.80	1.24
Colorado	1.48	2.33	0.89	1.21	1.58	2.45	0.59	1.24
Florida	1.40	2.22	0.77	1.17	1.60	2.83	1.15	1.26
Massachusetts	1.61	2.15	0.62	1.26	1.75	2.75	0.53	1.31
Michigan	1.47	2.58	0.75	1.20	1.78	2.89	0.54	1.31
Milwaukee	1.29	2.22	0.58	1.13	1.36	2.45	0.63	1.15
Minnesota	1.65	3.20	0.94	1.27	1.94	4.22	0.96	1.37
Mississippi	1.62	2.78	0.91	1.26	1.65	2.18	0.91	1.28
New Jersey	1.57	2.27	0.80	1.24	1.84	3.50	0.88	1.34
New York	1.46	2.44	0.75	1.20	1.73	2.59	0.59	1.31
Texas	1.38	2.24	0.34	1.16	1.65	3.24	0.74	1.27
Balance of the U.S.	1.44	2.28	0.85	1.19	1.49	2.20	0.65	1.21
Washington	1.39	2.43	1.00	1.17	1.51	2.69	0.84	1.22
National	3.81	5.32	2.02	1.94	3.91	5.85	2.83	1.97

Table 1-7
Average DEFF and DEFT for Estimates from the Child File for All
Children and Low-income Children, by Site

Study Area	All				Low-income			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.53	2.20	0.93	1.23	1.77	2.58	1.22	1.32
Balance of Wisconsin	1.47	2.41	0.74	1.20	1.78	3.20	1.10	1.32
California	1.85	3.11	0.89	1.35	2.32	7.04	0.86	1.49
Colorado	1.57	2.28	0.99	1.25	1.73	2.45	1.11	1.31
Florida	1.91	4.49	0.98	1.36	2.04	4.19	0.92	1.41
Massachusetts	1.69	2.87	0.90	1.29	2.12	3.61	1.28	1.45
Michigan	1.64	5.28	0.90	1.26	1.99	5.08	1.04	1.39
Milwaukee	1.63	2.83	0.72	1.27	2.03	2.80	1.22	1.42
Minnesota	2.65	14.14	1.19	1.56	2.88	14.48	1.20	1.62
Mississippi	1.99	4.79	0.88	1.39	2.24	4.83	1.06	1.47
New Jersey	1.60	2.77	0.89	1.25	1.99	3.79	1.11	1.40
New York	1.56	2.79	0.79	1.24	1.70	2.91	0.81	1.30
Texas	1.79	2.48	0.92	1.33	2.04	3.06	1.19	1.41
Balance of the U.S.	1.95	3.41	1.06	1.37	2.21	4.18	1.13	1.47
Washington	1.61	2.49	0.87	1.26	1.91	2.70	1.05	1.37
National	5.36	8.40	2.72	2.29	5.99	9.74	3.33	2.42

Table 1-8
Average DEFF and DEFT for Estimates from the Child File for
All Hispanic Children and Low-income Hispanic Children, by Site

Study Area	All				Low-income			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.50	2.58	0.58	1.21	1.57	3.04	0.56	1.22
Balance of Wisconsin	1.59	2.46	0.55	1.25	1.72	3.73	0.33	1.28
California	1.91	3.46	0.87	1.37	1.95	3.25	0.74	1.38
Colorado	1.64	2.36	1.01	1.27	1.67	2.37	0.91	1.28
Florida	1.47	2.39	0.87	1.20	1.57	2.33	1.03	1.25
Massachusetts	2.16	5.66	0.89	1.44	2.21	5.71	0.76	1.46
Michigan	1.58	2.76	0.58	1.24	1.62	3.42	0.71	1.26
Milwaukee	1.87	2.75	0.83	1.35	1.89	2.91	0.46	1.35
Minnesota	1.72	3.74	0.61	1.27	1.68	3.26	0.56	1.26
Mississippi	1.60	8.30	0.13	1.18	1.38	3.57	0.15	1.11
New Jersey	1.99	3.59	1.02	1.39	2.12	3.39	1.25	1.44
New York	1.56	2.14	0.88	1.24	1.55	2.21	0.95	1.24
Texas	1.99	3.15	1.09	1.40	2.03	3.64	0.93	1.41
Balance of the U.S.	1.62	2.80	0.63	1.26	1.70	3.29	0.79	1.29
Washington	1.89	3.43	0.75	1.36	2.01	3.83	0.83	1.40
National	3.54	5.31	1.94	1.87	3.54	5.89	1.70	1.86

Table 1-9
Average DEFF and DEFT for Estimates from the Child File for All
Black Children and Low-income Black Children, by Site

Study Area	All				Low-income			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.65	2.47	1.14	1.28	1.69	2.52	1.07	1.30
Balance of Wisconsin	1.73	4.59	0.38	1.26	2.01	4.56	0.44	1.36
California	1.87	3.82	0.88	1.35	1.90	4.20	0.77	1.34
Colorado	1.67	3.01	0.40	1.27	1.63	3.42	0.41	1.25
Florida	2.33	4.32	1.08	1.51	2.21	4.07	0.96	1.46
Massachusetts	1.93	4.64	1.02	1.37	1.70	3.23	1.12	1.29
Michigan	1.74	4.67	0.57	1.30	1.70	4.61	0.62	1.28
Milwaukee	1.92	3.35	1.38	1.38	2.15	3.37	1.19	1.46
Minnesota	2.62	11.77	0.67	1.52	2.55	11.79	0.36	1.49
Mississippi	2.30	5.19	1.15	1.49	2.30	5.26	0.74	1.48
New Jersey	1.75	3.14	1.10	1.31	1.85	3.62	0.79	1.34
New York	1.68	3.03	0.96	1.29	1.74	3.18	1.07	1.31
Texas	1.79	3.68	0.91	1.32	1.89	3.75	0.96	1.36
Balance of the U.S.	1.87	2.70	0.89	1.36	1.90	3.39	1.03	1.37
Washington	1.87	5.35	0.61	1.33	2.05	4.04	0.44	1.40
National	6.32	8.95	2.62	2.50	6.63	10.09	4.02	2.56

1.3 Structure of the Report

The previous section gave the results of computing and averaging a large number of design effects for both estimates of children and adults from the 1997 NSAF. These design effects are of primary interest to users of the data. They reveal that the design and estimation procedures used resulted in design effects that are significantly greater than what would be found in a simple random sample of the population. A simple random sample design was not even considered because it would have been prohibitively expensive and would not have achieved the sample sizes for the specific domains of interest in the NSAF, in particular for low-income households and for the 13 study areas. The design effects clearly indicate that the design and estimation procedures used in the survey need to be taken into account in the analysis of the data.

The design effects were computed using procedures that are described in the rest of the report. The next chapter briefly describes the replication and Taylor series approximation methods of computing sampling errors from complex samples. The third chapter goes further into why the replication method gives a more complete and accurate estimate of the sampling errors from the NSAF and describes how this method was applied. Chapter 3 also discusses some issues that arose in setting up the replicate structure. The fourth chapter shows how three popular software packages designed for computing sampling errors for complex samples can be used with the NSAF. The chapter also contains a discussion comparing the estimated sampling errors from the three packages. The final chapter summarizes the methods and findings associated with computing sampling errors in the survey. The appendices give examples of using the three software packages with the data files.

Other reports in this series are closely related to the material in this report. As discussed before, the Sample Design Report and part 1 of the Estimation Report series are essential because they provide the details on the sample design and estimation procedures that were used in the survey. These procedures determine how the sampling errors are computed. The Telephone and In-Person Survey Methods Reports are also closely related and give important information on how the survey was conducted.

Chapter 2

Methods for Variance Estimation

Variance estimation procedures have been developed to account for the sample design employed in a complex survey. Using these procedures, factors such as the selection of sample clusters in multi-stage sampling and the use of differential sampling rates to oversample a targeted subpopulation can be appropriately reflected in estimates of sampling error. The two main methods for estimating variances from a complex survey are known as replication and Taylor series estimation. Wolter (1985) is a useful reference on the theory and applications of these methods. The next two sections briefly review these methods.

2.1 Replication Methods

The basic idea behind replication is to draw subsamples from the sample, compute the estimate from each of the subsamples, and estimate the variance from the variability of the subsample estimates. Specifically, subsamples of the original “full” sample are selected to calculate subsample estimates of a parameter for which a “full-sample” estimate of interest has been generated. The variability of these subsample estimates about the estimate for the full sample can then be assessed. The subsamples are called replicates and the estimates from the subsamples are called replicate estimates. Balanced repeated replication (BRR) and jackknife replication are two general approaches to forming subsamples. Rust and Rao (1996) discuss these and other replication methods, show how the units included in the subsample can be defined using variance strata and units, and describe how these methods can be implemented using weights.

Replicate weights are created to derive the corresponding set of replicate estimates. Each replicate weight is derived using the same estimation steps as the full sample weight, but using only the subsample of cases comprising each replicate. Once the replicate weights are developed, it is a straightforward matter to compute estimates of variance for sample estimates of interest. Estimates of variance take the following form:

$$v(\hat{q}) = c \sum_{k=1}^G (\hat{q}_{(k)} - \hat{q})^2 \quad (2-1)$$

where

- \hat{q} is the estimate of q based on the full sample.
- $\hat{q}_{(k)}$ is the k -th estimate of q based on the observations included in the k -th replicate.
- G is the total number of replicates formed.
- c is a constant that depends on the replication method.
- $v(\hat{q})$ is the estimated variance of \hat{q} .

In the next chapter, the specific form of equation (2-1) used in the NSAF is presented. The use of WesVar Complex Samples (SPSS, 1998), a computer software program that generates

estimates of variance (standard errors, CVs, confidence intervals) with a specified set of replicate weights, is covered in chapter 4.

2.2 Taylor Series Method

The other widely used method for estimating variances in complex surveys is based on the Taylor series approximation. A Taylor series linearization of a statistic is formed and then substituted into the formula for calculating the variance of a linear estimate appropriate for the sample design. The Taylor series method relies on the simplicity associated with estimating the variance for a linear statistic even with a complex sample design. In most complex designs, the variance can be estimated by using the variance between PSUs and a replacement estimator (Wolter 1985). In this formulation, the strata and PSUs must be defined, similar to the variance estimation strata and units discussed above.

SUDAAN (1996) and STATA (1998) are two computer programs designed to produce variance estimates for complex surveys using the Taylor series method. Examples using these programs for the NSAF are given in the appendices for chapter 4. However, Taylor series was not the method chosen for producing most variance estimates from the survey, as discussed in the next chapter.

Chapter 3

Replication for the NSAF

This chapter has three sections. The first section describes why replication was chosen as the preferred method for computing variances from the survey data and why the particular form of the jackknife replication method was selected. The second section gives the details on setting up the replication structure, including the definition of the variance strata and units. The third section reviews some issues that arose during the development of this structure and provides some information about the consequences of handling those methods.

3.1 Choice of Method

The two major reasons for choosing replication as the primary method to estimate variances for the NSAF were operational convenience and the ability to reflect all components of the design and estimation in the estimates of variability. With respect to operational convenience, once replicate weights are constructed, it is very simple to compute estimates of sampling errors. No special care is needed for subgroups of interest, and no knowledge of the sample design is required. If an estimator is needed that was not previously considered, replication methods can be easily used to develop an appropriate estimate of variance. In such a case, variance estimates using a Taylor series approach would require additional work.

The second reason for choosing replication is probably more important. Variances are affected by both the nonresponse and poststratification types of adjustments made in developing the NSAF estimates. Replicate weights can be developed that reflect all such aspects of weighting. Currently, existing software for using the Taylor series method for variance estimation can only reflect the very last step in poststratification, and only then in specialized situations. Nonresponse adjustments and any other weight adjustments such as the raking described in *1997 NSAF Snapshot Survey Weighting* cannot be reflected in the variance estimates using the Taylor series method.

As mentioned in the previous chapter, the general idea behind the use of replication methods for estimating the variance of an estimate is to compare estimates of the same parameter generated from subsamples of the full sample to the estimate from the full sample. Sums of a function of the squared deviation of each subsample estimate from the corresponding full sample estimate provide estimates of sampling error. See equation (2-1).

Two different replication methods were considered for use in developing replicate weights; (1) a jackknife approach sometimes referred to as a random group methodology (JK1) and (2) a paired jackknife approach (JK2). For an RDD study, the JK1 method is often used since this method is appropriate for a single-stage unclustered approach with no explicit strata. The JK2 approach is more frequently employed where first-stage sampling units are selected with two units selected per explicit stratum. Thus, the JK2 approach would be appropriate for the area sample of PSUs.

The JK2 approach was used for both the RDD and area samples because it would be cumbersome to use two different replication methodologies in one survey. In the RDD sample, adjacent pairs of sampled telephone numbers were treated as having been sampled from the same

stratum (e.g., pairing the first and second numbers as sampled from variance stratum 1, the third and fourth numbers sampled from variance stratum 2, etc.). This was the approach used for another RDD study, the 1993 National Household Education Survey (Brick, et al, 1997).

The JK2 approach treats each pair of sampled telephone numbers as an implicit stratum, where each such stratum is defined by the sort order used in the sample selection of telephone numbers. It corresponds to the approach used for the area component of the study, the only difference being that explicit strata were constructed in the area sample prior to sample selection of PSUs.² In the JK2 method, the constant, c , in equation (2-1) is equal to 1. The next sections describe the details on the way replicates were prepared using the JK2 approach.

3.2 Design of Replicates

The first step in designing the replicate structure was to determine the number of variance estimation strata. In the JK2 method, the number of replicates is equal to the number of strata, so this really involves specifying the number of variance estimation strata. The choice of the number of variance estimation strata was based on the desire to obtain an adequate number of degrees of freedom to ensure stable estimates of variance while not having so many as to make the cost of computing variance estimates unnecessarily high. Generally, at least 30 degrees of freedom are needed to obtain relatively stable variance estimates. A number greater than 30 was targeted because there are other factors that reduce the contribution of a replicate to the total number of degrees of freedom, especially for estimates of subgroups. For example, estimating characteristics of the low-income population requires more strata because the low-income may be disproportionately distributed with a relatively large percentage in households without telephones.

A total of 60 variance estimation strata was chosen for the NSAF. The method of creating these is described below. An initial replicate structure with 82 variance estimation strata and replicates was implemented but was revised when the estimates for this scheme were inspected. The estimates were much higher than expected for a few estimates, and these were heavily influenced by the area sample. This topic is discussed more in the next section.

The variance estimation strata were created differently for the area and RDD samples. For the RDD component, a large number of variance estimation strata was possible since each pair of adjacent sampled phone numbers could be a variance stratum. Each telephone within the pair would constitute a variance unit. Such a large number of strata is unnecessary to achieve stable variance estimates. A total of 60 variance estimation strata were created for telephone numbers across all the geographical sites and for each site.

The variance strata for the RDD sample were formed as follows. First, the sampled telephone numbers were arranged in the same sort order used in sample selection. Next, adjacent sampled telephone numbers were paired to establish initial variance estimation strata (the first two sampled phone numbers were the first initial stratum, the third and fourth sampled telephone numbers were the second initial stratum, etc). Each telephone number in the pair was randomly

² The pairing of area sample PSUs for variance estimation is based on a "sort order" of the explicit strata since only one PSU was selected per stratum.

assigned to be either the first or second variance unit within the variance stratum. Each pair was sequentially assigned to one of 60 final variance estimation strata (the first pair to variance estimation stratum 1, the second to stratum 2, ..., the sixtieth pair to stratum 60, the sixty-first pair to stratum 1, etc.). As a result, each variance stratum had approximately the same number of telephone numbers. This was done by site.

In the original scheme, the variance estimation strata for the area sample were created separately for certainty or self-representing (SR) PSUs and noncertainty or nonself-representing (NSR) PSUs. This plan is described in DiGaetano, et al. (1998). In the revised replicate structure, the same process was applied to both types of PSUs. To create variance estimation strata, the segments were sorted and systematically paired in initial strata. The pairing of the segments was done within SR and NSR PSUs. Each segment was designated a variance unit. The initial variance strata were then combined to form 60 variance estimation strata, just as was done for the RDD sample. The 60 area variance estimation strata were then combined with the 60 variance strata from the RDD sample. The result was a total of 60 variance estimation strata, each containing some RDD and area sample units.

Once the variance strata were created, the replicate weights were created. The full replicate weights were constructed by modifying the full sample base weights. Replicate base weight for replicate k for record i was

$$\begin{aligned} w_i^{(k)} &= 2w_i \text{ if } i \text{ is in variance stratum } k \text{ and variance unit 1,} \\ &= 0 \text{ if } i \text{ is in variance stratum } k \text{ and variance unit 2,} \\ &= w_i \text{ if } i \text{ is not in variance stratum } k. \end{aligned}$$

The same sequence of weighting adjustments applied to the full sample weight was applied to the replicate base weight to create the final replicate weights. Thus, all of the different components of the weighting process were fully reflected in the replicate weights, ranging from household adjustments (nonresponse, adjustment for household noncoverage, and adjustment to control totals) to person adjustments (nonresponse and raking).

3.3 Issues Related to Replicate Structure

Analysis of important statistics based on the initial replication structure with 82 variance estimation strata showed large variances and design effects at the national and study area levels. It was suspected that the large variances were the result of unstable variance estimates due to the small number of NSR PSUs in the states and in the balance of the United States. The replicate structure was revised as discussed above largely to reduce the variance of the variance estimates. We opted to use the segments as the ultimate clusters (Hansen, Hurwitz, and Madow 1953) rather than the NSR PSUs. No changes were made to the variance estimation strata from the RDD sample and from the self-representing PSUs, where the segments were already the PSUs.

The main problem with the original replicate structure was that for a few important statistics a large contribution to the variance of the estimate was coming from the NSR PSUs. In most study areas and nationally, the number of NSR PSUs was very small — 4 to 6 in many sites — and the variance was very unstable due to this small number.

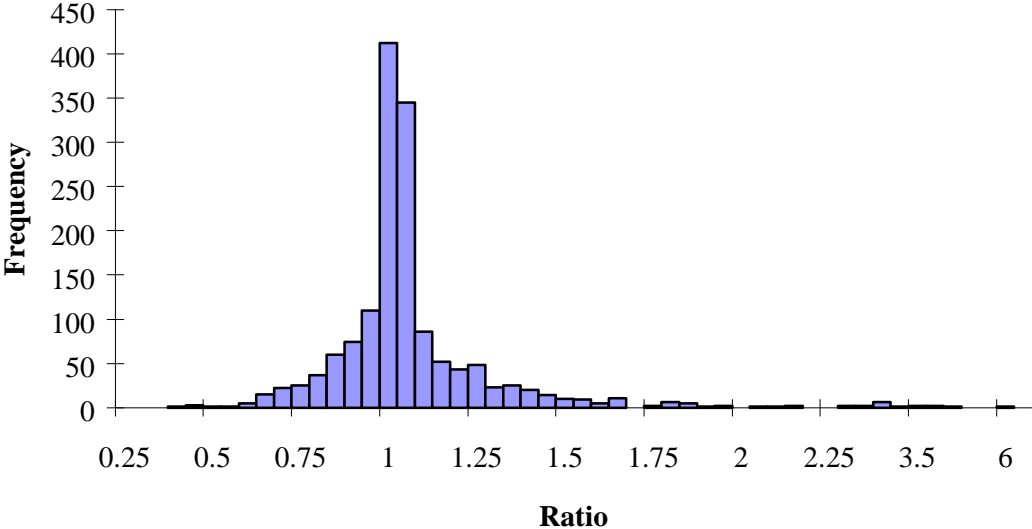
To address this problem, the approach suggested by Graubard and Korn (1996) was adapted. This approach essentially uses units selected at the second stage of sampling (segments in the case of the NSR PSUs) as the ultimate clusters rather than the first-stage units. The procedure thus ignores the between-PSU contribution of the NSR PSUs.

Before adopting this approach, two factors were considered. First, the original structure overestimated the variance of the estimates because two NSR PSUs had to be collapsed into one stratum for variance estimation purposes. Second, the effect of eliminating the NSR PSU contribution to the variance was estimated and found to be relatively small (less than 20 percent for estimates of nontelephone households). The main contribution to the variance was due to the sampling of segments and this was not affected by the change in the replicate structure.

To empirically assess the effects of the revision, replicate weights were created for both replicate structures. Variance estimates were then computed for a large number of statistics for site-specific variables (e.g., percent of children in Alabama who were not insured). The ratios of the variances under the original scheme to the revised scheme were examined.

Figure 3-1 shows a histogram of those ratios. The mean of the ratios is 1.048, the median is 1.00, and the first and third quartiles are 0.96 and 1.06, respectively. The variances are nearly the same under both schemes. The main effect of the revision was the reduction in the right skew of the distribution under the initial scheme. Even using the most severe and unrealistic criteria, the bias in the variance estimates in the revised scheme is less than 5 percent. Less severe criteria would suggest that the bias in the new scheme is small or trivial. We concluded the revised replicate structure provides more stable variance estimates with very small bias.

Figure 3-1
Ratio of DEFFs with Initial Variance Units to DEFFs with Revised Variance



Chapter 4

Software for Computing Variances

For complex sample surveys such as the NSAF, the computation of sampling errors requires specialized software. Many standard statistical software packages assume a simple random sample when computing estimates of variance. However, estimates of variance from these packages can seriously understate the true variability of the survey estimates. In recent years, specialized commercial software has been developed to analyze data from complex surveys (Lepkowski and Bowles 1996). The following sections describe the elements needed to compute estimates for the NSAF using three programs: WesVar Complex Samples, SUDAAN, and STATA.

4.1 WesVar

WesVar is a package developed by Westat and distributed by SPSS. WesVar uses replication methods to compute variance estimates. Through the use of replicates, adjustments made during weighting (nonresponse, raking) can be taken into account by making the same adjustments to each replicate separately. Replication is computer intensive, but powerful personal computers have largely eliminated this as an issue. However, it is still possible that for very large data sets the computations will exceed the capacity of the machine or take a long time. Although replication can be used for most estimates, replication techniques are not necessarily appropriate for all sample statistics of interest. Special care is needed when trying to estimate median, quartiles, or other quantiles. Direct estimates of quantiles using the jackknife method are not supported.

WesVar is an interactive program centered on sessions called “workbooks.” A workbook is a file linked to a specific WesVar dataset. In a workbook, the user can request descriptive statistics, as well as analyze and create new statistics. The information about the design is incorporated into the replicate weights when the data file is created. For descriptive statistics and analysis variables, WesVar offers the request. Regression requests support both linear and logistic requests. Outputs include statistics of interest, such as the sum of weights, means, percentages, along with their corresponding standard errors, design effects, coefficients of variation (CV), and confidence intervals. Appendix A shows an example of a request and the output using the child data file.

4.2 SUDAAN

SUDAAN (Software for the Statistical Analysis of Correlated Data) is a package developed by Research Triangle Institute (RTI) to analyze data from complex sample surveys. Like WesVar, SUDAAN computes standard errors of the estimates taking into account the survey design. SUDAAN and WesVar produce the same point estimates. The difference is in the method used to compute the variances. SUDAAN uses a first-order Taylor series approximation. As noted earlier, SUDAAN does not fully take account of complex weighting schemes such as nonresponse or raking. Medians and quantiles cannot be computed directly using this method because the functions do not satisfy the conditions needed for the approximation.

For descriptive statistics, SUDAAN offers two procedures: PROC CROSSTAB for categorical variables and PROC DESCRIPT for continuous variables. These procedures can be used to compute statistics of interest, such as sum of weights, means, and percentages along with their corresponding standard errors, design effects, and confidence intervals.

The weights developed for NSAF account for the dual-frame, complex sample design and include a complicated weighting process with nonresponse and raking adjustments. The first aspect is handled by SUDAAN with the use of additional variables that account for the sample design. The creation of these variables is described in appendix B. SUDAAN does not have any procedure to handle raking. When a sample is raked, the variance of an estimate is reduced since the totals are known without sampling variation. Using SUDAAN without any modifications will produce standard errors of estimates that do not reflect this reduction in variance due to the raking. Note that this is not an issue for replication, where the effect of raking can be included when the replicate weights are created.

The estimates of the standard errors can be improved by using SUDAAN's poststratification option. This option reflects the reduction in variance due to adjusting to control totals in one dimension. However, this approach still does not reflect the full effect of raking since the other raking dimensions are ignored. There are also some restrictions in using the poststratification option. The option is valid only in PROC DESCRIPT and design effects are not computed with this option. While using PROC DESCRIPT with poststratification for continuous variables is very simple, additional instructions are required to display the percentages for all the levels of categorical variables. This makes the code awkward to use.

To reflect some of the effects of raking in the estimates, we "re-poststratified" the weights to the sum of weights in the same cells used in one of the raking dimensions. If there were no missing records or missing values of the analytical variables, then the sum of weights would be the same as the cell control totals used in raking to this dimension. When the weights are re-poststratified in this way, the point estimate of the statistics do not change since the weights were already adjusted to the control totals. However, the standard errors will be smaller and closer to the ones produced by WesVar. It is important to note that, although the approach is valid, the way it is used is not common. This procedure is valid as long as the analytical values have no missing values or the file contains all the records used in the weighting process. Small departures are acceptable, but the procedure is more complex if there are many missing values.

An example of computing estimates in SUDAAN is given in appendix B. The first part of the appendix gives the code needed to reflect the sample design. The next part shows the program and statistics produced using PROC CROSSTAB.

The third part of the appendix shows an example on how to run SUDAAN with the poststratification option. The process requires creating a variable in the file that indicates the cells to which the weight is being poststratified. Then control totals are computed by adding the weights of the nonmissing values of the variable being analyzed. If the variables have very different missing value patterns, the control totals should be computed for each variable and in separate SUDAAN runs. If there are a few missing values or missing records, this may not be necessary but will result in small errors in the point estimates.

4.3 STATA

STATA is a command driven, fully programmable statistical package used for managing, analyzing, and graphing data. STATA was developed by StataCorp and is available for a variety of platforms, including DOS, Windows, Macintosh, and UNIX. STATA statistical, graphical, and data management capabilities are fully expandable through programming.

There is a family of commands in STATA developed to analyze data from sample surveys. Using these commands, STATA can compute standard errors of estimates taking the survey design into account, but only for a few procedures such as descriptive statistics and regression analysis. Like SUDAAN, STATA uses the Taylor series method. The information about the design is incorporated through the use of auxiliary variables that indicate the strata and PSUs. In this way, the effect of stratification and clustering is reflected in the standard errors of the estimates. Besides point estimates (proportions, means, and totals) and their standard errors, STATA can compute confidence intervals, design effects, and misspecification effects. Design and misspecification effects are computed for means only.

Although STATA is an improvement over general-purpose statistical software, it cannot properly reflect the reductions of variance due to complex weighting schemes such as nonresponse, poststratification, or raking. However, the main difficulty with using STATA for the NSAF is that it requires at least two PSUs in each variance stratum before any statistics can be computed. If there are not two or more PSUs per stratum, the user has to manually collapse the strata before the program runs. Furthermore, since missing values are excluded from the data, different collapsing may be needed for analytical variables with different patterns of missing values. STATA does not have a readily available option to overcome this restriction like SUDAAN does.

Like SUDAAN, STATA requires auxiliary variables to define the strata and PSUs. The same variables that reflect the original sample design created to analyze the data in SUDAAN (see appendix B) can be used for STATA. However, these variables do not meet the condition of having at least two PSUs per stratum. The first part of appendix C describes how the file can be manipulated so that STATA can be run.

Three STATA commands used to analyze survey data are **svytotal**, **svyprop** and **svymean**. These are used to estimate totals, proportions, and means, respectively. The command **svyprop** does not produce DEFFs for proportions. The command **svymean** can be used to produce the DEFFs for proportions by coding the analytical variable with values 0 and 1. For totals, a variable ONE must be created with a value of 1 for all the records. Examples of the use of these commands are given in the second part of appendix C.

4.4 Comparison of WesVar, SUDAAN, and STATA

The three software programs for estimating sampling errors from complex samples were used to compute sampling errors from the 1997 NSAF child file to illustrate some of the issues discussed above. Table 4-1 has the output from each of the programs for six different variables. As noted

previously, all three programs produce the same estimate and are based on the same sample size, so the first columns in the tables are not repeated for each program. The remaining columns (columns 4 through 11) do vary from program to program and they are examined below.

Columns 4 and 5 have the estimated standard errors and DEFFs from WesVar for the six statistics. For example, the estimated standard error for the percent of children who live in families at or below 100 percent of the poverty level (F100POV) is .534 and the DEFF for this estimate is 6.279. No DEFFs are produced for estimates of totals.

Columns 6 and 7 have the corresponding estimated standard error and DEFF estimates producing using SUDAAN without using the poststratification option. The standard errors and DEFFs are much greater than those produced by WesVar. For example, the DEFF for the percent of children who live in families at or below 100 percent of the poverty level from SUDAAN is 8.33 rather than the 6.279 from WesVar. Almost all of this difference is because SUDAAN, without the poststratification option, does not account for the adjustment to the control totals. This is immediately obvious by looking at the standard errors of the totals (labeled marginal in the table), where the WesVar estimate is zero and the SUDAAN estimate is very large. We return to this point after discussing the STATA estimates.

Columns 10 and 11 have the STATA estimates using the same data. These estimates are most similar to the SUDAAN results in columns 6 and 7 because they also do not account for the adjustment to the control totals. The only reason that the SUDAAN and STATA estimates of standard errors and DEFFs are not equal is because the programs handle strata with only one PSU differently. In STATA, the user must manually collapse any such strata, which generally results in slight overestimates of the standard errors. SUDAAN does not collapse the strata but uses a different algorithm for strata with only one PSU. The algorithm is relatively complex and described in the SUDAAN manuals, but it often results in overestimates of the variance. As can be seen from the values in the table, it appears that in these cases the overestimates of the standard errors from the SUDAAN algorithm are larger than those associated with the collapsing done in STATA. In any event, both the SUDAAN and STATA naïve estimates are substantial overestimates because they do not take the adjustment to control totals into account.

Columns 8 and 9 give the estimated standard errors from SUDAAN when the poststratification option is used as described in appendix B. The estimates in column 8 result from the poststratification to the study area/race/ethnicity/age/sex dimension, while those in column 9 result from the poststratification to a home ownership/region distribution (selected study areas and remaining census regions). As can be seen from these estimates, the poststratification effect brings the estimates of standard errors (SUDAAN does not produce DEFFs with this option) into line with the WesVar estimates. The estimates are not exactly equal for several reasons. The three most important reasons are: (1) The methods of variance estimation are different and these affect the estimates; (2) SUDAAN reflects only one of the two dimensions used in the person-level adjustment process while WesVar includes both; and (3) The WesVar estimates include the effect of nonresponse and household-level poststratification while the SUDAAN estimates do not.

The estimates in the table show that the estimates from the 1997 NSAF can be produced using any of the three software packages, but the WesVar outputs are more appropriate because they do account for most of the important features of the design and estimation process. The SUDAAN package with the poststratification option provides reasonable estimates. Because STATA cannot handle adjustments to control totals, it is less useful than the other two. All three packages are much more appropriate than general purpose software such as SAS and SPSS.

Table 4-1
Estimates of Children Computed Using WesVar, SUDAAN, and STATA

Variable	All Software			WesVar		SUDAAN		SUDAAN Poststratification (1stDimension)	SUDAAN Poststratification (2ndDimension)	STATA	
	EST_TYPE	N	ESTIMATE	STDERROR	DEFF	STDERROR	DEFF	STDERROR	STDERROR	STDERROR	DEFF
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
F100POV											
1	VALUE	7,329	13,527,354	371,936		487,472		392,594	363,324	475,778	
0	VALUE	27,110	56,090,868	371,936		884,084		392,594	363,324	882,539	
MARGINAL	VALUE	34,439	69,618,222	0		1,010,354			0		
1	PERCENT	7,329	19.43	0.534	6.279	0.615	8.33	0.564	0.522	0.611	8.208
0	PERCENT	27,110	80.57	0.534	6.279	0.615	8.33	0.564	0.522	0.611	8.208
MARGINAL	PERCENT	34,439	100.00			0				0	
F200POV											
1	VALUE	17,027	29,024,768	360,460		631,696		474,273	431,643	597,660	
0	VALUE	17,412	40,593,454	360,460		829,001		474,273	431,643	828,938	
MARGINAL	VALUE	34,439	69,618,222	0		1,010,354			0		
1	PERCENT	17,027	41.69	0.518	3.798	0.748	7.922	0.681	0.620	0.729	7.522
0	PERCENT	17,412	58.31	0.518	3.798	0.748	7.922	0.681	0.620	0.729	7.522
MARGINAL	PERCENT	34,439	100.00			0			0	0	
UNINS											
1	VALUE	4,547	8,038,628	239,475		275,699		252,797	255,736	272,508	
0	VALUE	29,892	61,579,594	239,475		954,235		252,797	255,736	938,235	
MARGINAL	VALUE	34,439	69,618,222	0		1,010,354			0		
1	PERCENT	4,547	11.55	0.344	3.99	0.375	4.739	0.363	0.367	0.373	4.702
0	PERCENT	29,892	88.45	0.344	3.99	0.375	4.739	0.363	0.367	0.371	4.702
MARGINAL	PERCENT	34,439	100.00			0			0	0	
JAFDCY											
1	VALUE	30,930	62,771,675	311,769		360,500		304,990	300,898	357,073	
0	VALUE	3,509	6,846,547	311,769		938,189		304,990	300,898	925,401	
MARGINAL	VALUE	34,439	69,618,222	0		1,010,354			0		
1	PERCENT	30,930	90.17	0.448	7.789	0.483	9.073	0.438	0.432	0.482	9.022
0	PERCENT	3,509	9.83	0.448	7.789	0.483	9.073	0.438	0.432	0.482	9.022
MARGINAL	PERCENT	34,439	100.00			0			0	0	

Table 4-1
Estimates of Children Computed Using WesVar, SUDAAN, and STATA (Continued)

Variable	All Software			WesVar		SUDAAN		SUDAAN Poststratification (1stDimension)	SUDAAN Poststratification (2ndDimension)	STATA	
	EST_TYPE	N	ESTIMATE	STDERROR	DEFF	STDERROR	DEFF	STDERROR	STDERROR	STDERROR	DEFF
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
KLUNCH											
1	VALUE	10,107	17,472,060	338,148		496,480		386,713	379,918	478,335	
2	VALUE	6,777	11,502,609	301,140		320,996		327,071	325,389	317,453	
-1	VALUE	17,245	39,988,369	380,472		853,333		486,694	474,163	852,851	
MARGINAL	VALUE	34,129	68,963,038	84,530		998,613			0		
1	PERCENT	10,107	25.34	0.486	4.266	0.630	7.981	0.562	0.551	0.619	6.916
2	PERCENT	6,777	16.68	0.439	4.743	0.472	7.157	0.473	0.471	0.471	5.439
-1	PERCENT	17,245	57.99	0.544	4.151	0.756	5.452	0.700	0.685	0.747	7.809
MARGINAL	PERCENT	34,129	100.00			0			0	0	
JFSTAMPY											
1	VALUE	27,985	57,473,079	407,645		486,180		401,763	381,788	469,682	
0	VALUE	6,454	12,145,143	407,645		909,480		401,763	381,788	905,853	
MARGINAL	VALUE	34,439	69,618,222	0		1,010,354			0		
1	PERCENT	27,985	82.55	0.59	8.199	0.630	9.494	0.577	0.548	0.617	9.109
0	PERCENT	6,454	17.45	0.59	8.199	0.630	9.494	0.577	0.548	0.617	9.109
MARGINAL	PERCENT	34,439	100.00			0			0	0	

Chapter 5

Summary

This report presents the methods for computing estimates of sampling errors from the 1997 NSAF, along with a summary of some of these measures. In the first chapter, tables of average design effects are given for both estimates of children and adults. These estimates show that design procedures, especially the differential sampling rates for low-income households and the concentration of the sample in the 13 focal states, resulted in design effects that were significantly greater than one. For example, the design effect for all adults for national estimates is about 4.7, while within each state the average design effect is less than half that amount.

The methods used to produce these design effects and to produce direct estimates of sampling errors for other statistics are presented, as well. A jackknife replication method was chosen as the primary method of computing sampling errors from the 1997 NSAF, primarily because this method can easily accommodate the nonresponse and raking adjustments used in the estimation process.

Both the replication and Taylor series approximation method of computing sampling errors require the definition of variance estimation strata and units. The report describes how the instability of the variance estimators that arose due to the small number of clusters sampled in the area sample was overcome. By using sampled segments as the variance units rather than the NSR PSUs, the stability of the variance estimates was greatly improved while keeping the bias of the variance estimators very small.

Rather than relying on average design effects, sampling errors can be computed directly from the NSAF data files. Three different software packages that can be used to produce direct estimates of sampling error are described. Procedures for running WesVar, the software appropriate for replicated sampling errors, are given. These procedures are the ones used to compute the average design effects tables in chapter 1. The procedures for the Taylor series software, SUDAAN and STATA, are also presented. SUDAAN is easier to use than STATA for this survey because it has an automatic procedure for handling situations with only one PSU per stratum. Examples of all three software runs and outputs are given in the appendices.

References

Brick, J. M., E. Tubbs, M. Collins, and M. J. Nolin, 1997. *Unit and Item Response, Weighting and Imputation Procedures in the 1993 National Household Education Survey*, Working Paper 97-05. U.S. Department of Education, Office of Educational Research and Improvement.

DiGaetano, R., J. M. Brick, and I. Flores-Cervantes, 1998. *Preserving Degrees of Freedom in the Multi-Mode, Multi-Site Survey*. In Proceedings of the Survey Research Methods Section of the American Statistical Association.

Flores-Cervantes, I., G. Shapiro, and S. Brock-Roth, 1998. *Effect of Oversampling by Poverty Status in an RDD Survey*. In Proceedings of the Survey Research Methods Section of the American Statistical Association.

Graubard, B., and E. Korn, 1996. *Modeling the Sampling Design in the Analysis of Health Surveys*. *Statistical Methods in Medical Research*, 5: 263–281.

Hansen, M. W. Hurwitz, and W. Madow, 1953. *Sample Survey Methods and Theory*, vols. I and II. John Wiley and Sons.

Kish, L. 1985. *Survey Sampling*. John Wiley and Sons.

Kish, L. 1992. *Weighting for Unequal Pi*, *Journal of Official Statistics*, 8: 183–200.

Lepkowski, J., and J. Bowles, 1996. *Sampling Error Software for Personal Computers*. *The Survey Statistician*, 35: 10–16.

Rust, K.F. and J. N. K. Rao, 1996. *Variance Estimation for Complex Surveys Using Replication Techniques*. *Statistical Methods in Medical Research*, 5: 282–310.

Skinner, C.J., D. Holt, and T. M. F. Smith, 1989. *Analysis of Complex Surveys*. John Wiley and Sons.

SPSS (1998), *WesVar® Complex Samples™ 3.0*, User's Guide, SPSS.

STATA (1998), *User's Guide*, STATA.

SUDAAN (1996), *SUDAAN User's Manual*, Release 7.0. Research Triangle Institute.

Verma, V., C. Scott, and C. O'Muircheartaigh, 1980. *Sample Designs and Sampling Errors for the World Fertility Survey*. *Journal of the Royal Statistical Society, A* 143: 431–473.

Waksberg, J., J. M. Brick, G. Shapiro, I. Flores-Cervantes, and B. Bell 1997. *Dual-frame RDD and Area Sample for Household Survey with Particular Focus on Low-income Population*. In Proceedings of the Survey Research Methods Section of the American Statistical Association, 713-718.

Wolter, Kirk M. 1985. *Introduction to Variance Estimation*. Springer-Verlag.

Appendix A WesVar Example

The following section describes how to compute sample estimates and their corresponding standard errors using WesVar. In the example, proportions, totals, and their standard errors are computed for the variable UNINS (Focal child currently uninsured) from the Child data file.

Required Information

The information about the design is already incorporated in the replicate weights. The variance estimation method (JK2) to be used is provided during the creation of the WesVar data file. Unlike SUDAAN or STATA, WesVar does not require explicit variables to indicate the variance strata and sampled PSUs. This information was used during the creation of the replicate weights but it is not needed to compute sample estimates. The NSAF data files have one full sample final weight and 60 final replicate weights. The output of a workbook used to estimate for UNINS is shown below.

WesVar Output

Summary Information of Table Request One

TIME THE JOB EXECUTED:	13: 17: 34 3/17/1999
INPUT DATASET NAME:	C: \ui \Ch. var
TIME THE INPUT DATASET CREATED:	15: 24: 29 03/15/1999
OPTION SUMMARY IS:	ON
OPTION FUNCTION LOG IS:	ON
OPTION SPELLING LABEL IS:	OFF
OPTION VALUE LABEL IS:	ON
OPTION OUTPUT REPLICATE ESTIMATES IS:	OFF
VARIANCE ESTIMATION METHOD:	JK2
FINITE POPULATION CORRECTION FACTOR:	1. 00000
VALUE OF ALPHA (CONFIDENCE INTERVAL %):	0. 05000 (95. 00000 %)
DEGREES OF FREEDOM	60
t VALUE:	2. 000
OPTION COMPLETE IS:	ON
FULL SAMPLE WEIGHT:	CHA5W0
REPLICATE WEIGHTS:	CHA5W1. . . CHA5W60
ANALYSIS VARIABLES:	None Specified.
COMPUTED STATISTIC:	None Specified.
TABLE(S):	UNINS
FACTOR(S):	1. 00
NUMBER OF REPLICATES:	60
NUMBER OF OBSERVATIONS READ:	34439
WEIGHTED NUMBER OF OBSERVATIONS READ:	69618222. 000

TABLE : UNINS

UNINS	STATISTIC	EST_TYPE	ESTIMATE	STDERROR	CV(%)	N	DEFB
0	SUM_WTS	VALUE	61579594.22	239475.430	0.389	29892	N/A
1	SUM_WTS	VALUE	8038627.78	239475.430	2.979	4547	N/A
MARGINAL	SUM_WTS	VALUE	69618222.00	0.000	0.000	34439	N/A
0	SUM_WTS	PERCENT	88.45	0.344	0.389	29892	3.990
1	SUM_WTS	PERCENT	11.55	0.344	2.979	4547	3.990
MARGINAL	SUM_WTS	PERCENT	100.00	.	.	34439	.

*Warning: One dimensional table. Rowpct and colpct options ignored.

Appendix B

SUDAAN Example

The following section describes how to compute sample estimates and their corresponding standard errors using SUDAAN. The first example produces estimates and standard errors without using the poststratification option. In the second example, the same estimates are computed using the poststratification to one dimension as an approximation for raking (two dimensions). The examples use the variable UNINS (Focal child currently uninsured) from the Child data file.

Required Variables

To reflect the effect of the design in the variance estimation, SUDAAN requires variables that indicate the variance estimation strata and sampled PSUs. The variable for the variance strata (STRATA) is created as follows:

```
STRATA=1000*SITE + 100*FLAG+NVARSTRT
```

Where FLAG is defined as:

$$\text{FLAG} = \begin{cases} 0 & \text{If SMPTYPE} = \text{"A"} \\ 1 & \text{If SMPTYPE} = \text{"R"} \end{cases}$$

The variable that defines the selected PSU within strata is NVARUNIT. These two variables should be used in the NEST statement as:

```
NEST STRATA NVARUNIT / STRLEV=1 PSULEV=2 MISSUNIT.
```

The keyword MISSUNIT is required to compute the contribution to the variance in strata with only one PSU.

For categorical variables, SUDAAN requires positive integer values. All categorical variables being analyzed should be recoded to meet this requirement. This means any variables with '0' and '1' need to be recoded to be '1' and '2.'

The keyword that reflects the design is WR. For additional descriptions of the options and keywords, refer to the SUDAAN manual.

Special attention is required when the sequential variable (PSTCELL) that indicates the poststratification cell is created. For the first dimension, PSTCELL is created by sequentially numbering the levels of the cross-tabulation of SITE * CHA3CL (for children) or SITE * AD1A3CL (for adults). For the second dimension, PSTCELL is created by sequentially numbering the levels of cross-tabulation of UIREG*TENURE (for children) or UIREG*EDUC*TENURE (for adults).

UIREG is defined as follows:

$$\text{UIREG} = \begin{cases} \text{SITE} & \text{If SITE in (2,3,8,10) \quad \text{These are CA, FL, NY, TX}} \\ 100 & \text{otherwise} \end{cases}$$

Once the cell definition is created in the data file, compute the control total for the cell (POSTWGT) by adding CHA5W0 for children or AD1A5W0 for adults. The totals are hardcoded in the program in the same order indicated by the cell PSTCELL. Since PROC DESCRIPT was designed to handle continuous variables, the procedure presents the results in a strange format when categorical variables are analyzed.

Running SUDAAN Without Poststratification

```

                S U D A A N
      Software for the Statistical Analysis of Correlated Data
      Copyright      Research Triangle Institute      May 1998
                Release 7.5.2
*****
This copy of Stand-Alone Network SUDAAN, serial number P0000079, for Windows
is licensed to WESTAT, Inc. (The SUDAAN Administrator).
*****
1  PROC CROSSTAB DATA="d:\uisudan\newchild"
      FILETYPE=SAS DESIGN=WR DEFT ; 2  WEIGHT CHA5W0 ;
3  NEST STRATA NVARUNIT / STRLEV=1 PSULEV=2 MISSUNIT ;
4  SUBGROUP  UNINS
      ;
5  LEVELS  2  ;
6  SETENV DECWDTH=3 COLWIDTH=13 LINESIZE=80;
7  TABLE UNINS ;
8  PRINT NSUM WSUM SEWGT DEFFWGT ROWPER SEROW DEFFROW
      TOTPER SETOT DEFFTOT /STYLE=NCHS ;

      /*print nsum wsum sewgt deffwgt rowper serow deffrow
      /style=nchs ; */
```

Opened SAS data file d:\uisudan\newchild.SSD for reading.

DATA WARNING in Request 30:

There is a problem with nest variable STRATA=1001.000000 in record 2
It has only one NVARUNIT whose value is 2.000000
Standard fixup is to use the square of the taylorized deviation for
NVARUNIT=2.000000 as the contribution to the variance.

(more warnings)

DATA WARNING in Request 64:

There is a problem with nest variable STRATA=1024.000000 in record 48
It has only one NVARUNIT whose value is 1.000000
Standard fixup is to use the square of the taylorized deviation for
NVARUNIT=1.000000 as the contribution to the variance.

Date: 03-18-99
Time: 12:29:47

Research Triangle Institute
The CROSSTAB Procedure

Page : 1
Table : 1

Variance Estimation Method: Taylor Series (WR)
by: FC currently uninsured.

FC currently uninsured	Sample Size	Weighted Size	SE Weighted
Total	34439.000	69618222.000	1010353.563
1	4547.000	8038627.779	275699.488
2	29892.000	61579594.221	954234.745

Date: 03-18-99
Time: 12:29:47

Research Triangle Institute
The CROSSTAB Procedure

Page : 2
Table : 1

Variance Estimation Method: Taylor Series (WR)
by: FC currently uninsured.

FC currently uninsured	Row Percent	SE Row Percent	DEFF Row Percent #4
Total	100.000	0.000	.
1	11.547	0.375	4.739
2	88.453	0.375	4.739

Variance Estimation Method: Taylor Series (WR)
by: FC currently uninsured.

FC currently uninsured	Tot Percent	SE Tot Percent	DEFF Tot Percent #4
Total	100.000	0.000	.
1	11.547	0.375	4.739
2	88.453	0.375	4.739

456 strata or clusters at a given stage of the design with only one subunit.
Variance contribution for that strata or cluster is computed using the
deviation from the overall mean of that sampling stage.

228 strata or clusters at a given stage of the design with only one subunit.
Variance contribution for that strata or cluster is computed using the
deviation from the overall mean of that sampling stage.

228 strata or clusters at a given stage of the design with only one subunit.
Variance contribution for that strata or cluster is computed using the
deviation from the overall mean of that sampling stage.

CROSSTAB used

CPU time : 42.0 seconds
Elapsed time : 43 seconds
Virtual memory : 1.95 MB

Running SUDAAN With Poststratification

S U D A A N

Software for the Statistical Analysis of Correlated Data
Copyright Research Triangle Institute May 1998
Release 7.5.2

This copy of Stand-Alone Network SUDAAN, serial number P0000079, for Windows
is licensed to WESTAT, Inc. (The SUDAAN Administrator).

```
1 PROC DESCRIPT DATA="d:\ui\sudan\newchild"  
   FILETYPE=SAS DESIGN=WR ;  
2 WEIGHT CHA5W0 ;  
3 NEST STRATA NVARUNIT / STRLEV=1 PSULEV=2 MISSUNIT ;  
  
4 SUBGROUP _ONE_ PSTCELL UNINS ;  
5 LEVELS 1 332 2 ;  
6 POSTVAR PSTCELL ;  
7 POSTWGT  
   12308.00226 28398.003919 30332.005429 38130.003966  
   38023.00058 42002.000111 27956.001845 29470.006474 37092.008999  
   37740.001167 41469.003505 60737.000753 62249.996529 80013.994749  
   81363.997387 84975.992934 56944.995819 58539.99761 75700.994207  
   77197.99062 81144.993732 368045.0833 368027.14288 392358.08114  
   344903.09944 339351.07147 350352.11278 352648.07617 375131.07436  
   328965.0239 309652.98923 201504.09137 219493.0847 216090.06433  
   418925.99725 461533.91541 559487.9136 526544.82304 482861.89186  
   398527.916 438415.98126 530011.88699 499889.8704 460129.80914  
   50429.05421 49195.061605 60182.042632 116736.05316 47193.993856  
   46472.97781 57407.982183 109855.0172 58094.113334 66076.082224  
   86109.077003 83272.060518 76783.083924 57023.088552 64154.078368  
   84380.090881 80632.050897 77045.046232 183858.05278 195445.93175  
   260732.91084 250403.88456 232834.83029 174369.04582 185215.97151  
   246729.92316 238239.75505 222381.73964 26345.008057 15253.00461  
   25553.00747 25121.00741 14427.004622 24709.008425 41385.010659  
   31054.005946 30894.005914 94515.001369 103153.00014 146511.00004  
   136910.00097 128890.99287 90097.002824 98027.998177 138395.99761  
   130135.99108 124987.99158 34635.00635 28403.999674 27638.005296  
   75204.019732 52714.02004 46008.005541 45404.003574 73382.009884  
   50841.009427 44936.000317 45599.999634 157864.98779 168470.98215  
   232693.97934 237048.9723 238205.96769 150442.98742 160318.99096  
   220823.9834 225036.97225 227137.96595 12654.001089 18702.00262  
   16474.005056 15791.003978 15663.003748 89474.996797 94531.990909  
   136474.98433 143629.98535 143328.97717 85712.993209 90300.995918  
   129652.98904 136597.98145 136676.98107 28252.006972 28346.008887  
   32343.006303 29266.003857 29974.006406 26952.00708 26661.00581  
   30827.007535 55870.012298 26528.010353 28907.008813 37130.011046  
   34134.008164 33927.010227 25323.003596 28394.005087 35774.008377  
   32854.007411 33392.005259 117049.99495 125892.99514 165067.97967
```

154039. 97915 146799. 9853 111828. 99153 120067. 99336 156901. 99079
146085. 98853 138712. 99343 86328. 124346 84327. 11016 89906. 096147
79805. 080188 80632. 098437 81620. 097696 80473. 086568 86338. 119536
77710. 085075 78237. 06241 58871. 080508 66643. 042515 93415. 102866
84957. 016835 81907. 071265 57200. 086944 64793. 075028 90949. 103803
83028. 104994 82629. 078822 252062. 84921 272847. 91187 347191. 81585
326299. 70165 313961. 67501 239845. 89782 259171. 92958 328681. 89378
309745. 78901 297143. 81207 209868. 11828 193208. 15913 215889. 04327
212187. 99279 216190. 04536 199043. 15382 185368. 11218 206851. 09793
203924. 037 204106. 98506 113717. 13845 242379. 15842 110243. 15649
159373. 02527 78750. 061105 233853. 03731 245023. 95986 318530. 93329
322776. 80101 312763. 698 223357. 90402 233708. 90889 303017. 81965
307956. 79745 303169. 85596 25833. 006226 14304. 00359 26608. 001897
24261. 006962 13473. 003647 24368. 003015 15889. 003167 19798. 004077
19055. 005436 101605. 01056 108285. 9968 145943. 99653 150298. 99818
147508. 99159 96896. 008043 103200. 00023 138917. 99711 142287. 99216
141715. 98845 6369. 9998727 27912. 002056 29333. 004313 36654. 00408
38677. 00443 42004. 002581 27671. 004632 28607. 00205 35664. 002708
37679. 0038 41310. 999992 33446. 998056 34812. 997219 45116. 995661
47396. 995205 49735. 995742 31636. 000956 32923. 999654 42658. 997093
44519. 99636 47460. 993636 4265. 0013908 6902. 001099 4086. 0011722
6858. 0015198 6717. 0025393 7395. 0022929 9793. 0027803 9277. 0028967
8733. 0012956 6650. 0023843 7132. 0025329 9307. 0034476 8938. 002312
8879. 00279 11543. 999874 11287. 999554 15145. 999963 15673. 998853
15748. 998485 10963. 000559 10738. 99936 14456. 998898 14962. 000348
15342. 998953 14129. 002285 13086. 000307 21846. 003398 77720. 997435
83963. 996654 121371. 99256 131514. 98898 133057. 98545 73910. 996001
79841. 994091 115400. 99448 123788. 9877 126725. 98772 425543. 0898
237023. 05557 443951. 00315 404067. 09655 645768. 08572 945333. 26024
1843292. 239 439862. 10732 482396. 10339 620674. 0629 1191573. 2378
2156347. 9382 2255529. 0163 3019334. 8243 3074987. 769 3102962. 701
2049017. 0889 2142390. 9874 2865599. 9295 2917512. 8356 2959955. 7758
13944. 003921 15936. 003298 15788. 003868 18837. 002083 18054. 001866
21879. 001932 21274. 002242 20927. 001649 18107. 002883 17526. 002398
20729. 001753 20012. 002354 19679. 001768 61640. 000856 64990. 999847
88815. 995392 92236. 99421 89009. 994261 58789. 998809 61823. 99791
84067. 995824 87342. 992494 84665. 992519

;

8 SETENV DECWIDTH=3 COLWIDTH=12 LINESIZE=80;

9 VAR UNINS UNINS ;

10 CATLEVEL 1 2 ;

11 TABLE _ONE_ ;

12 PRINT NSUM WSUM TOTAL SETOTAL PERCENT SEPERCENT

/STYLE=NCHS ;

Opened SAS data file d:\uisudan\newchild.SSD for reading.

Number of observations read : 34439 Weighted count : 69618222

Denominator degrees of freedom : 994

DATA WARNING in Request 71:

There is a problem with nest variable STRATA=1001.000000 in record 2
It has only one NVARUNIT whose value is 2.000000
Standard fixup is to use the square of the taylorized deviation for
NVARUNIT=2.000000 as the contribution to the variance.

DATA WARNING in Request 94:

There is a problem with nest variable STRATA=1024.000000 in record 48
It has only one NVARUNIT whose value is 1.000000
Standard fixup is to use the square of the taylorized deviation for
NVARUNIT=1.000000 as the contribution to the variance.

Date: 03-19-99
Time: 07:53:44

Research Triangle Institute
The DESCRIPT Procedure

Page : 1
Table : 1

Variance Estimation Method: Taylor Series (WR)
Post-stratified estimates
by: Variable, One.

Variable One	Sample Size	Weighted Size	Total

FC currently uninsured: 1			
Total	34439.000	69618222.000	8038627.779
1	34439.000	69618222.000	8038627.779
FC currently uninsured: 2			
Total	34439.000	69618222.000	61579594.221
1	34439.000	69618222.000	61579594.221

Date: 03-19-99
Time: 07:53:44

Research Triangle Institute
The DESCRIPT Procedure

Page : 2
Table : 1

Variance Estimation Method: Taylor Series (WR)
Post-stratified estimates
by: Variable, One.

Variable			
One	SE Total	Percent	SE Percent
<hr/>			
FC currently			
uninsured: 1			
Total	252796.661	11.547	0.363
1	252796.661	11.547	0.363
FC currently			
uninsured: 2			
Total	252796.661	88.453	0.363
1	252796.661	88.453	0.363

228 strata or clusters at a given stage of the design with only one subunit.
Variance contribution for that strata or cluster is computed using the
deviation from the overall mean of that sampling stage.
228 strata or clusters at a given stage of the design with only one subunit.
Variance contribution for that strata or cluster is computed using the
deviation from the overall mean of that sampling stage.

DESCRIPT used
CPU time : 36.0 seconds
Elapsed time : 36 seconds
Virtual memory : 2.27 MB

Appendix C STATA Example

The following section describes how to compute sample estimates and their corresponding standard errors using STATA. There is no poststratification option in STATA. In the first example, proportion and totals and their standard errors are computed for the variable UNINS (Focal child currently uninsured) from the Child data file. In the second example, additional issues arising from the presence of missing data are addressed. The second example uses the variable KLUNCH (Free lunches at school last year) from the same file.

Required Variables

To reflect the effect of the design in the variance estimation, STATA requires variables for the variance estimation strata and for the sampled PSUs. For purposes of these examples, these variables can be created as follows:

$$\text{STRATA2} = \begin{cases} \text{See Attachment 1} & \text{If the value of STRATA is found in Table C - 1} \\ \text{STRATA} & \text{Otherwise} \end{cases}$$
$$\text{NVARUNT2} = \begin{cases} \text{See Attachment 1} & \text{If the value of STRATA2 is found in Table C - 1} \\ \text{NVARUNIT} & \text{Otherwise} \end{cases}$$

It is important to note that if STATA is used to produce statistics, the variables STRATA2 and NVARUNT2 should be created more carefully. We need to take into account how the original PSUs were collapsed when the sample was drawn so that the new collapsed strata do not cross area/telephone samples and sites. The additional collapsing of the single PSU strata was carried out in three steps. The incomplete strata were sorted sequentially and grouped in pairs. The new stratum then corresponds to the smallest original stratum within the pair. If the value of the PSUs (NVARUNT2) was the same for the two strata, the first one was modified. The collapsing did not cross sample type or site.

The variables STRATA2 and NVARUNT2 can be used to compute estimates for all analytical variables with no missing values. If there are missing values, additional collapsing may be needed. This is shown in the last section of this appendix when estimates are computed for KLUNCH.

To define the design in STATA, we used the following commands:

<u>Commands</u>	<u>Description</u>
svyset strata strata2	Define strata
svyset psu nvarunt2	Define PSU
svyset pweight cha5w0	Define weight

Another useful instruction is “**svydes variable**” that verifies that in the data there are no single PSU strata after dropping the records with missing values of the variable being analyzed.

Computing Estimates for KLUNCH

The variable KLUNCH has missing values. Although the strata were collapsed, the presence of missing values creates two additional single PSU strata. We can use the command **svydesc** to determine which strata are incomplete. Using other STATA commands we can create new variables STRATA3 and NVARUNT3 that reflect the additional collapsing. Now we can proceed to compute the statistics in the same way we did for the other variables. Note that if there are a lot of analytical variables with different patterns of missing variables, then the computation of the statistics can be very slow and cumbersome. Also, note that additional work is needed to compute the design effect for proportions and for variables with more than two levels not coded with values of 0 and 1.

Table C-1
Strata with only one nvarunit

OBS	STRATA	NVARUNIT	STRATA2	NVARUNT2
1	1001	2	1001	1
2	1002	2	1001	2
3	1003	2	1003	1
4	1004	2	1003	2
5	1005	2	1005	2
6	1007	1	1005	1
7	1008	2	1008	2
8	1014	1	1008	1
9	1019	1	1019	2
10	1024	1	1019	1
11	1025	1	1025	2
12	1028	1	1025	1
13	1029	1	1029	1
14	1030	2	1029	2
15	1031	2	1031	2
16	1032	1	1031	1
17	1035	1	1035	1
18	1040	2	1035	2
19	1042	2	1042	1
20	1047	2	1042	2
21	1049	1	1049	1
22	1050	2	1049	2
23	1051	1	1051	1
24	1052	2	1051	2
25	2002	2	2002	2
26	2003	1	2002	1
27	2030	2	2030	1
28	2031	2	2030	2
29	2035	2	2035	1
30	2036	2	2035	2
31	2049	2	2049	1
32	2051	2	2049	2
33	2053	1	2053	1
34	2058	1	2053	1
35	2060	2	2053	2
36	3004	1	3004	1
37	3009	2	3004	2
38	3011	1	3011	2
39	3013	1	3011	1
40	3020	1	3011	1
41	3025	2	3025	2
42	3029	1	3025	1
43	3030	2	3030	2
44	3032	1	3030	1
45	3033	1	3033	1
46	3034	2	3033	2
47	3041	1	3041	2
48	3042	1	3041	1
49	3043	1	3043	2
50	3044	1	3043	1
51	3045	2	3045	2
52	3050	1	3045	1
53	3053	1	3053	1
54	3054	3	3053	3
55	4003	2	4003	1
56	4010	2	4003	2
57	4014	2	4014	1
58	4017	2	4014	2
59	4021	1	4021	1
60	4022	2	4021	2
61	4028	2	4028	1
62	4029	2	4028	2
63	4031	1	4031	1
64	4058	2	4031	2
65	5003	2	5003	1
66	5033	2	5003	2

Table C-1
Strata with only one nvarunit (continued)

OBS	STRATA	NVARUNIT	STRATA2	NVARUNT2
67	5034	1	5034	1
68	5036	2	5034	2
69	5039	2	5039	1
70	5043	2	5039	2
71	5047	1	5047	2
72	5051	1	5047	1
73	5052	2	5052	1
74	5055	2	5052	2
75	6001	1	6001	1
76	6003	2	6001	2
77	6033	1	6033	1
78	6038	2	6033	2
79	6041	2	6041	1
80	6043	2	6041	2
81	6045	2	6045	2
82	6047	1	6045	1
83	6058	2	6058	2
84	6059	1	6058	1
85	7002	2	7002	1
86	7008	1	7008	1
87	7010	2	7008	2
88	7013	2	7013	1
89	7015	2	7013	2
90	7016	1	7016	2
91	7020	1	7016	1
92	7023	1	7023	2
93	7024	1	7023	1
94	7026	2	7026	1
95	7027	2	7026	2
96	7028	1	7028	1
97	7037	2	7028	2
98	7038	1	7038	1
99	7041	2	7038	2
100	7045	2	7045	2
101	7046	1	7045	1
102	7051	1	7051	2
103	7053	1	7051	1
104	7054	2	7054	2
105	7056	1	7054	1
106	7058	1	7058	1
107	7059	2	7058	2
108	8008	1	8008	2
109	8013	1	8008	1
110	8014	2	8014	1
111	8015	2	8014	2
112	8020	1	8020	1
113	8030	2	8020	2
114	8032	2	8032	1
115	8033	2	8032	2
116	8034	2	8034	1
117	8035	2	8034	2
118	8036	1	8036	2
119	8039	1	8036	1
120	8042	1	8042	1
121	8044	2	8042	2
122	8045	1	8045	1
123	8056	2	8045	2
124	8059	2	8045	2
125	10007	1	10007	2
126	10012	1	10007	1
127	10017	1	10017	2
128	10019	1	10017	1
129	10021	1	10021	2
130	10022	1	10021	1
131	10023	1	10023	2
132	10024	1	10023	1

Table C-1
Strata with only one nvarunit (continued)

OBS	STRATA	NVARUNIT	STRATA2	NVARUNT2
133	10025	2	10025	1
134	10026	2	10025	2
135	10028	2	10028	1
136	10052	1	10052	1
137	10055	2	10052	2
138	10057	1	10057	2
139	10059	1	10057	1
140	10060	1	10057	1
141	11015	1	11015	2
142	11017	1	11015	1
143	11019	1	11019	2
144	11021	1	11019	1
145	11023	2	11023	2
146	11048	1	11023	1
147	11052	2	11023	2
148	13005	1	13005	2
149	13008	1	13005	1
150	13011	2	13011	2
151	13012	1	13011	1
152	13013	2	13013	1
153	13014	2	13013	2
154	13017	1	13017	1
155	13018	2	13017	2
156	13020	1	13020	1
157	13021	2	13020	2
158	13022	2	13022	1
159	13023	2	13022	2
160	13025	2	13025	2
161	13027	1	13025	1
162	13028	2	13028	2
163	13032	1	13028	1
164	13036	2	13036	2
165	13040	1	13036	1
166	13041	2	13041	1
167	13042	2	13041	2
168	13045	1	13041	1
169	14005	2	14005	1
170	14006	2	14005	2
171	14008	1	14008	2
172	14009	1	14008	1
173	14012	1	14012	1
174	14014	2	14012	2
175	14015	1	14015	1
176	14017	2	14015	2
177	14020	1	14020	2
178	14021	1	14020	1
179	14022	2	14022	2
180	14026	1	14022	1
181	14027	2	14027	1
182	14029	2	14027	2
183	15006	2	15006	2
184	15008	1	15006	1
185	15009	1	15009	2
186	15014	1	15009	1
187	15018	2	15018	1
188	15021	2	15018	2
189	15023	1	15022	1
190	15026	2	15022	2
191	15027	1	15022	1
192	16003	2	16003	1
193	16004	2	16003	2
194	16005	2	16005	2
195	16006	1	16005	1
196	16007	3	16007	1
197	16013	2	16007	2
198	16019	1	16019	1

Table C-1
Strata with only one nvarunit (continued)

OBS	STRATA	NVARUNIT	STRATA2	NVARUNT2
199	16020	2	16019	2
200	16021	1	16021	2
201	16022	1	16021	1
202	16025	2	16025	1
203	16037	2	16037	2
204	16038	1	16037	1
205	16040	1	16040	1
206	16043	2	16040	2
207	16045	1	16045	1
208	16049	2	16045	2
209	16050	1	16050	2
210	16051	1	16050	1
211	16053	2	16053	2
212	16059	1	16053	1
213	16060	2	16053	2
214	18004	2	18004	2
215	18005	1	18004	1
216	18007	1	18007	2
217	18010	1	18007	1
218	18011	1	18011	1
219	18012	2	18011	2
220	18019	1	18019	1
221	18020	2	18019	2
222	18021	1	18021	2
223	18022	1	18021	1
224	18026	2	18026	2
225	18041	1	18026	1
226	18048	2	18048	2
227	18050	2	18048	2
228	18060	1	18048	1

Example of a STATA Session

```
set memory 32000
(32000k)
. insheet using newchild.txt, comma
(9 vars, 34439 obs)
```

```
. svyset strata strata2
. svyset psu nvarunt2
. svyset pweight cha5w0
```

```
. gen one=1
```

```
. svytotal one, by ( unins )
```

Survey total estimation

```
pweight: cha5w0      Number of obs   =   34439
Strata:   strata2    Number of strata =   1100
PSU:     nvarunt2    Number of PSUs  =   2204
                          Population size = 69618222
```

Total	Subpop.	Estimate	Std. Err.	[95% Conf. Interval]	Deff
one					
	unins==0	61579594	938234.6	5.97e+07 6.34e+07	61.24098
	unins==1	8038628	272507.9	7503936 8573320	5.166273

```
. svyprop unins
```

```
pweight: cha5w0      Number of obs   =   34439
Strata:   strata2    Number of strata =   1100
PSU:     nvarunt2    Number of PSUs  =   2204
                          Population size = 69618222
```

Survey proportions estimation

unins	_Obs	_EstProp	_StdErr
0	29892	0.884533	0.003734
1	4547	0.115467	0.003734

```
. svymean unins
```

Survey mean estimation

```
pweight: cha5w0      Number of obs   =   34439
Strata:   strata2    Number of strata =   1100
PSU:     nvarunt2    Number of PSUs  =   2204
                          Population size = 69618222
```

Mean	Estimate	Std. Err.	[95% Conf. Interval]	Deff
unins	.1154673	.0037343	.1081401 .1227945	4.702109

```
. svytotal one, by ( klunch)
stratum with only one PSU detected
r(460);
```

```
. svydes klunch
```

```
pweight:  cha5w0
Strata:    strata2
PSU:      nvarunt2
```

Strata strata2	#PSUs included	#PSUs omitted	#Obs with complete data	#Obs with missing data	#Obs per included PSU		
					mi n	mean	max
1001	2	0	3	0	1	1.5	2
1003	2	0	6	0	2	3.0	4

```
(more output)
```

3043	2	0	3	0	1	1.5	2
3045	1*	1	2	1	2	2.0	2
3053	2	0	31	0	1	15.5	30
3101	2	0	26	0	7	13.0	19

```
(more output)
```

16044	2	0	6	0	2	3.0	4
16045	1*	1	1	1	1	1.0	1
16047	2	0	4	0	2	2.0	2

```
(more output)
```

18160	2	0	38	0	16	19.0	22
1100	2202	2	34129	310	1	15.5	42

			34439				

```
. gen strata3=strata2
```

```
. replace strata3=3053 if strata2==3045
(3 real changes made)
```

```
. replace strata3=16047 if strata2==16045
(2 real changes made)
```

```
. gen nvarunt3= nvarunt2
```

```
. svyset psu strata3
```

```
. svyset strata strata3
```

```
. svyset psu nvarunt3
```

```
. svyset pweight cha5w0
```

```
. svytotal one, by ( klunch)
```

Survey total estimation

```
pweight:  cha5w0          Number of obs   =   34129
Strata:   strata3        Number of strata =   1098
PSU:     nvarunt3        Number of PSUs  =   2201
                               Population size = 68963038
```

```
-----+-----
```

Total	Subpop.	Estimate	Std. Err.	[95% Conf. Interval]	Deff
-----+-----					
one					
	klunch== -1	39988369	852850.5	3.83e+07 4.17e+07	21.42419
	klunch== 1	17472060	478334.6	1.65e+07 1.84e+07	8.67959
	klunch== 2	11502609	317453.2	1.09e+07 1.21e+07	5.203613

```
-----+-----
```

. svyprop klunch

```
-----+-----
```

pweight:	cha5w0	Number of obs	=	34129
Strata:	strata3	Number of strata	=	1098
PSU:	nvarunt3	Number of PSUs	=	2201
		Population size	=	68963038

```
-----+-----
```

Survey proportions estimation

klunch	_Obs	_EstProp	_StdErr
-1	17245	0.579852	0.007466
1	10107	0.253354	0.006192
2	6777	0.166794	0.004706

. gen klunch1=klunch

(310 missing values generated)

. replace klunch1=0 if klunch== -1

(17245 real changes made)

. replace klunch1=0 if klunch== 2

(6777 real changes made)

. gen klunch2=klunch

(310 missing values generated)

. replace klunch2=0 if klunch== 1

(10107 real changes made)

. replace klunch2=0 if klunch== -1

(17245 real changes made)

. replace klunch2=1 if klunch== 2

(6777 real changes made)

. gen klunch_1=klunch

(310 missing values generated)

```
. replace klunch_1=1 if klunch==- 1
(17245 real changes made)
```

```
. replace klunch_1=0 if klunch==1
(10107 real changes made)
```

```
. replace klunch_1=0 if klunch==2
(6777 real changes made)
```

```
. table klunch1 klunch
```

```
-----+-----
      |          KLUNCH
klunch1 |   - 1      1      2
-----+-----
      0 | 17245          6777
      1 |          10107
-----+-----
```

```
. table klunch2 klunch
```

```
-----+-----
      |          KLUNCH
klunch2 |   - 1      1      2
-----+-----
      0 | 17245  10107
      1 |          6777
-----+-----
```

```
. table klunch_1 klunch
```

```
-----+-----
      |          KLUNCH
klunch_1 |   - 1      1      2
-----+-----
      0 |          10107  6777
      1 | 17245
-----+-----
```

```
. svymean klunch1
```

Survey mean estimation

```
pweight:  cha5w0          Number of obs   =   34129
Strata:   strata3        Number of strata =   1098
PSU:     nvarunt3        Number of PSUs  =   2201
                          Population size = 68963038
```

```
-----+-----
      Mean | Estimate  Std. Err.  [95% Conf. Interval]      Deff
-----+-----
klunch1 | .253354   .0061915   .2412055   .2655025   6.916175
-----+-----
```

```
. svymean klunch2
```

Survey mean estimation

```
pweight: cha5w0      Number of obs   =   34129
Strata:   strata3    Number of strata =   1098
PSU:     nvarunt3    Number of PSUs  =   2201
                        Population size = 68963038
```

Mean	Estimate	Std. Err.	[95% Conf. Interval]		Deff
klunch2	.1667938	.0047061	.15756	.1760277	5.438669

. svymean klunch_1

Survey mean estimation

```
pweight: cha5w0      Number of obs   =   34129
Strata:   strata3    Number of strata =   1098
PSU:     nvarunt3    Number of PSUs  =   2201
                        Population size = 68963038
```

Mean	Estimate	Std. Err.	[95% Conf. Interval]		Deff
klunch_1	.5798522	.0074661	.5652029	.5945015	7.808635
