A Comparison of Programs for Variance Estimation from Complex Sample Surveys

Wayne E. Johnson, Lester R. Curtin, Van L. Parsons
National Center for Health Statistics

This poster reports results from work conducted in the Office of Research and Methodology; the aim was to become familiar with several statistical packages in order to provide advice and support for users in the Center.

Parts of four packages were employed:

(1) the DESCRIPT procedure in **SUDAAN** from Research Triangle Institute,
(2) the SURVEYMEANS procedure in **SAS,**
(3) the SVYMEAN procedure in **STATA**, and
(4) **WESVAR4** from WESTAT.

These packages were used to produce estimates for the means and standard errors of three variables for adults sampled in a national survey of households: the 1997 National Health Interview Survey. The variables were HEIGHT and WEIGHT, which were recorded for most, but not all, respondents, and the AGE OF FIRST CANCER - - the age of onset of the first occurrence of cancer from among thirty specific types, such as blood, bladder, bone, brain, etc. - - AGE OF FIRST CANCER was available for only a small proportion of the respondents.

The public use file for the 1997 National Health Interview Survey is available on CD-ROM, Series 10, Number 12A. This work used the ASCII version issued December 1999.

The initial sampling weight was taken to be the inverse of the probability of selection; this was adjusted for case nonresponse; weights were further adjusted through post-stratification in order to match national totals. The actual survey design has been approximated with a simplified public-use-file design: two PSU's drawn with replacement from each of 339 strata.

A few words about each of the packages: SUDAAN has been used at the Center for many years as a production tool - - it is SAS-callable and current versions run on personal computers and UNIX workstations. A legacy version on the IBM mainframe still receives heavy use. Some of our newest employees have been educated in schools of public health where they have already received training in the use of the STATA package, which runs on personal computers and UNIX workstations. STATA runs fast, in part because it stores datafiles in RAM memory. The SAS Institute has developed its own procedures for analyzing survey data: SURVEYMEANS and SURVEYREG; these are part of the SAS/STAT package which is available on all the platforms and operating systems supported by the SAS Institute. Each of these first three packages can be used interactively or in batch mode; the fourth, WESVAR4, is designed to be used in an interactive fashion alone. WESVAR4 runs on personal computers and UNIX work stations, and, reflecting its origin within a firm which specializes in statistical design and consulting, incorporates unique capabilities for adjusting sampling weights (WESVAR4 will even develop balanced repeated replicate weights).

From the descriptions of the packages' algorithms, one would expect close agreement among the standard errors produced by SAS's SURVEYMEANS, STATA's SVYMEAN and SUDAAN's DESCRIPT for the stipulated design: simple random sampling with replacement within strata; one would not expect exact agreement between their standard errors and those produced by WESVAR4, because WESVAR4 uses a different method of variance estimation. As expected, all four packages produced the same estimates for the mean values themselves; to avoid repetition, these are displayed only once, rather than over and over for

each package. For the same reason, economy of presentation, it was decided to choose one of the first three packages for complete, detailed display of its standard errors; the others' are shown either as "=" if in agreement or, if different, are also shown in detail (the arbitrary degree of precision for "agreement" is: no difference through the fourth decimal place - the ten-thousandths). At the Center, SUDAAN is the most heavily used of these packages; as the package to be so distinguished, it was a natural choice - no other token of "favor" is intended.

STATA detected this condition: there were strata in which observations on AGE OF FIRST CANCER occurred in only one of the PSU's, rather than both. In this situation, STATA advises the user to consider collapsing strata and dismisses the request for a table. The authors considered that users of the NHIS public use file would have no basis for deciding how to collapse strata (because actual design information is deliberately obscured for reasons of confidentiality). The authors chose a method of collapsing strata and found that STATA ran and gave reasonable estimates, but feel that these ought not to be included here because of subjectivity. For AGE OF FIRST CANCER, STATA's column for standard errors in Table 3 contains "n/a".

As Tables 1-3 show, there are discrepancies among the estimated standard errors which would not have been expected. The discrepancies appear to be related to subdomain size - - getting worse as the number of observations becomes smaller.

Comparison of Standard Errors for Mean Height

Sampled Adults, 1997 National Health Interview Survey, NCHS CD-ROM Series 10, No. 12A

Table 1				Height	(inches)				
		N	NMISS	MEAN	STD ERR SUDAAN WR & JK	STD ERR STATA	STD ERR SAS	STD ERR WESVAR JKN	STD ERR WESVAR BRR
REGION	NorthEast	6653	450	66.7278	0.0481	=	=	=	=
	MidWest	7830	500	67.1374	0.0478	=	=	=	=
	South	12068	784	67.0465	0.0424	=	=	=	=
	West	7282	549	66.8766	0.0598	=	=	=	0.0599
ETHNICITY	Hispanic	5239	953	65.5927	0.0572	=	0.0571	=	0.0575
	Non-Hispanic	28576	1834	67.1208	0.0261	=	=	=	=
RACE	White	27210	1647	67.0722	0.0263	=	=	=	=
	Black	4801	722	66.8944	0.0696	=	0.0694	=	=
	Other	1822	903	65.4592	0.1036	=	0.1035	=	0.1041
SEX	Male	14545	952	69.9645	0.0268	=	=	=	0.0267
	Female	19288	1331	64.2118	0.0192	=	=	=	=

‘ = ’ signifies that the cell’s contents agreed with those of the corresponding cell for SUDAAN to at least the fourth decimal place.

SUDAAN WR & JK: SUDAAN’s DESCRIPT procedure with design = WR | JK for With Replacement and Jackknife, respectively. Their results agreed at least to the fourth decimal place.

STATA: STATA’s SVYMEAN procedure.

SAS: SAS’s PROC SURVEYMEANS.

WESVAR JKN & BRR: WESVAR’s Jackknife N and Balanced Repeated Replications options.

Comparison of Standard Errors for Mean Weight

Sampled Adults, 1997 National Health Interview Survey, NCHS CD-ROM Series 10, No. 12A

Table 2				Weight (pounds)						
		N	NMISS	MEAN	STD ERR SUDAAN WR & JK	STD ERR STATA	STD ERR SAS	STD ERR WESVAR JKN	STD ERR WESVAR BRR	
REGION	NorthEast	6489	614	165.6266	0.5493	=	=	=	0.5482	
	MidWest	7705	625	168.6597	0.5478	=	=	=	0.5479	
	South	11842	1010	167.6983	0.3699	=	=	=	0.3702	
	West	7184	647	164.2916	0.5675	=	=	=	0.5691	
ETHNICITY	Hispanic	5193	1116	162.3945	0.5548	=	0.5527	=	0.5571	
	Non-Hispanic	28009	2401	167.3742	0.2645	=	=	=	0.2644	
RACE	White	26716	2141	166.7342	0.2575	=	=	=	=	
	Black	4706	855	173.8359	0.6503	=	0.6482	=	0.65	
	Other	1798	989	153.7227	1.0155	=	1.0141	1.0156	1.0231	
SEX	Male	14473	1024	185.0635	0.3325	=	=	=	0.3326	
	Female	18747	1872	149.6592	0.2688	=	=	=	0.2689	

‘ = ’ signifies that the cell’s contents agreed with those of the corresponding cell for SUDAAN to at least the fourth decimal place.

SUDAAN WR & JK: SUDAAN’s DESCRIPT procedure with design = WR | JK for With Replacement and Jackknife, respectively. Their results agreed at least to the fourth decimal place.

STATA: STATA’s SVYMEAN procedure.

SAS: SAS’s PROC SURVEYMEANS.

WESVAR JKN & BRR: WESVAR’s Jackknife N and Balanced Repeated Replications options.

Comparison of Standard Errors for Mean Age of First Cancer

(Age at onset was collected for thirty specific types of cancer; Age of First Cancer is the minimum of these for each patient)

Sampled Adults, 1997 National Health Interview Survey, NCHS CD-ROM Series 10, No. 12A

Table 3 Age of Onset for First Cancer Among Thirty Categories of Cancer

		N	NMISS	MEAN	STD ERR SUDAAN WR	STD ERR SUDAAN JK	STD ERR STATA	STD ERR SAS	STD ERR WESVAR JKN	STD ERR WESVAR BRR
REGION	NorthEast	424	6679	53.5782	1.1118	=	n/a	1.1029	=	1.1223
	MidWest	536	7794	51.8483	0.7762	=	n/a	0.7509	=	0.7777
	South	861	11991	51.2198	0.5723	=	n/a	0.5613	=	0.5746
	West	535	7296	49.3149	0.8716	=	n/a	0.8582	=	0.8816
ETHNICITY	Hispanic	137	5548	45.6758	2.3484	2.3504	n/a	1.2392	2.3504	2.3973
	Non-Hispanic	2219	28191	51.635	0.3972	=	n/a	0.387	=	0.398
RACE	White	2119	26738	51.743	0.412	=	n/a	0.4014	=	0.4139
	Black	182	5026	49.4861	1.3608	1.361	n/a	1.0883	1.361	1.3636
	Other	55	1996	43.4339	2.6203	2.6266	n/a	0.7225	2.6266	2.6882
SEX	Male	843	14654	56.3502	0.609	=	n/a	0.5339	=	0.6117
	Female	1513	19106	48.0036	0.5183	=	n/a	0.4979	=	0.5177

‘ = ’ signifies that the cell’s contents agreed with those of the corresponding cell for SUDAAN’s WR to at least the fourth decimal place.

SUDAAN WR & JK: SUDAAN’s DESCRIPT procedure with design = WR | JK for With Replacement and Jackknife, respectively.

STATA: There were strata in which observations belonged to only one PSU; user required to collapse such strata.

SAS: SAS’s PROC SURVEYMEANS.

WESVAR JKN & BRR: WESVAR’s Jackknife N and Balanced Repeated Replications options.

Sample Distribution for Height, Weight and Age of First Cancer

Sample Adults, 1997 National Health Interview Survey, NCHS CD-ROM Serioes 10, No. 12

Table 4		Height			Weight			Age 1st Cancer		
		None	One	Two	None	One	Two	None	One	Two
ALL		0	0	339	0	0	339	2	46	291
REGION	NorthEast	273	0	66	273	0	66	273	8	58
	MidWest	260	0	79	260	0	79	260	13	66
	South	210	4	125	210	4	125	217	19	103
	West	264	0	75	264		75	265	10	64
ETHNICITY	Hispanic	55	82	202	57	80	202	262	62	15
	Non-Hispanic	0	0	339	0	0	339	9	55	275
RACE	White	1	0	338	1	0	338	10	55	274
	Black	43	78	218	43	79	217	237	80	22
	Other	59	96	184	59	97	183	293	46	0
SEX	Male	0	0	339	0	0	339	42	129	168
	Female	0	0	339	0	0	339	13	76	250

Cells display the number of strata in the sample which contain {none, one, or two} PSU's with observations belonging to the subdomain (row) for the variables Height, Weight and Age of First Cancer. Height and Weight were collected for most subjects; Age of First Cancer was undefined for most subjects.