Alcohol and Drug Services Study (ADSS), 1996-1999: [United States]

United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Office of Applied Studies

Codebook for Part 11: Phase II/III Stratified Jackknife Factors

1. Published statistics, including a few variables in this codebook, may not be exactly reproducible from the data in the public use file due to the disclosure protection procedures that were implemented.

2. The Data File User's Manuals provided in the codebooks contain references to SAS transport databases originally created by the data producers. To provide the data to users in a format that is neither system nor platform specific, the data files are in ASCII text format with SAS and SPSS data definition statements. Additionally, the number of variables found in the data files differ from the original number of variables cited by the data producers. The unweighted frequencies provided in the codebooks correspond to the data files.

3. In the Client Abstract data files, for the variable A62, "TEST RESULTS", the abstractor's instructions were to code "1 = Positive (leave blank if negative or not applicable)". Accordingly, negative test results were combined with inapplicable responses that are coded as -9. Any analysis of this series will be affected by this combining of negative and inapplicable responses.

4. In the Client Abstract data files, a new variable was created for A65 by the data producers: "TREATMENT EPISODES IN THE LAST 12 MONTHS". Therefore the questionnaire and variable information do not match. The new variable provides the number of treatment episodes in the prior 12 months, rather than a dichotomous response to whether or not the respondent had any treatment during this timeframe.

5. Disclosure analysis was performed on the ADSS files by SAMHDA, resulting in modifications to the data. These are explained in the following section, "Confidentiality Protection".

6. The Phase I facility public use file includes 2394 of the original 2395 records. One facility's record was deleted due to the presence of outlying data that could potentially identify the facility.

7. The Stratified Jackknife Factor files for Phase I and Phase II/III list values for the jackknife replication factors for use with the SUDAAN and WesVar statistical software only. These files are not intended for use with other statistical packages.

8. The Stratified Jackknife Factor files are space-delimited ASCII data files containing 1 record each. A detailed description of the use of these files is included in this codebook.

9. The jackknife factors are in the order expected by WesVar. The first factor corresponds to the first replicate, the second corresponds to the second replicate, and so on to the 200th factor, which corresponds to the 200th replicate.

10. The Phase I Finite Population Correction file contains the finite population correction factors (FPC) for use with the WesVar and SUDAAN statistical software only. The space-delimited ASCII data file contains 200 records and 1 variable. This file is not intended for use with other statistical packages.

11. The FPCs are in the order expected by WesVar. The first FPC corresponds to replicate 1, the second FPC to replicate 2, and so on to the 200th FPC, which corresponds to the 200th replicate.

Confidentiality Protection

Disclosure analysis for the ADSS files was conducted by the Substance Abuse and Mental Health Data Archive (SAMHDA). Measures taken to protect the confidentiality of the ADSS facility and client records included (1) using microaggregation for problematic variables, (2) deleting direct identifier variables such as facility name, and (3) recoding variables. The disclosure protection procedures allow nearly all of the data to be publicly released, take into consideration the most likely analytic uses of the data, and ensure the confidentiality of both facilities and clients.

Microaggregation

Microaggregation as applied to ADSS involved identifying problematic variables, sorting records by the first problematic variable, grouping records into three based on their value for this variable, averaging the values for each grouping, and applying the average to the records in each group. This was repeated for each of the problematic variables, which included the client count and financial data found in the Phase I Facility File. Cells with values of zero were excluded from microaggregation.

Microaggregation is a recoding method in which each variable has a set of ranges defined for it. For each variable, the range replaces each true record value. Such ranges (recodes) are usually defined summarily, irrespective of the data; in microaggregation the data themselves determine the ranges. The values most impacted by this approach are likely to be outliers or the values at either tail of a distribution. In other types of disclosure procedures, however, those values would be suppressed or top- or bottom-coded, which typically distorts the data substantially more than microaggregation (e.g., $500,000; $678,000; and $1,750,000 would become "$500,000 or more"). Microaggregation was preferable to these other methods because it allows statistics such as measures of central tendency to be run (e.g., to obtain average client counts and revenues), which are likely to be of interest to researchers. Researchers may want to categorize the ADSS data in performing their own analyses. Microaggregation allows them to do this in whatever way works best for them, without attempting to pre-determine the categories that would work for the most analysts.

The steps involved in the microaggregation were to:

1.   Identify the problematic variables.
2.   Microaggregate the variables identified, excluding values of zero.
3.   Recalculate variables as necessary, based on the variables that were microaggregated.

Two Phase I variables were microaggregated: total substance abuse treatment revenue (D7) and total clients in all types of care on October 1, 1996 (B1J2). The total treatment revenue (D7) was carried forward to two additional variables (D8TOT and D12D). All of these "total revenue" variables provided the same data and respondents were instructed to copy the D7 total to D8TOT and D12D. All three of these variables were treated as microaggregated variables in determining the impact to the data.

The microaggregated variables were included in tables in the facility questionnaire that specified breakdowns of total revenue and client counts (the B1, B2 and D8 tables). Therefore, it was necessary to address the problem of having columns within the tables add correctly. Each cell within these tables represents a different variable. The totals were microaggregated and the number in each cell was recalculated by applying the relative percentage of the total for each cell. Totals were microaggregated, rather than sub-parts of the tables because all records had totals but not all records had valid numbers in the other cells in the tables. The more records that are microaggregated, the more closely the records are likely to cluster and the less impact there is to the data. These tables included 191 variables.

The only change to the Phase II Administrator file was the carrying over of the total substance abuse treatment revenue value from Phase I. This is Q52 in the Phase II file. No changes were made to the client files, other than the deletion of administrative variables and variables such as date of birth.

The ADSS Cost Study included a computerized desk audit to check for consistency and accuracy of data previously collected in the Phase I Facility and Phase II Administrator files. Three post-audit Cost Study variables were microaggregated: NB12, ND7, and NQ52. Related variables were recalculated based on the microaggregated data.

Results of Microaggregation

In order to assess the impact to the data, for the microaggregated and recalculated variables, the cells that changed more than five percent in either direction were calculated as a percentage of valid cells (including zero) and as a percentage of total cells. Because a large number of valid values in the data are zero, we also calculated the cells that changed more than five percent as a percentage of non-missing and non-zero cells. We included all three revenue variables as microaggregated, though the original values for all three variables were the same. The results are provided in Table 1 and show that less than one percent of the non-missing and non-zero microaggregated variables changed more than five percent, while 3.6 percent of the recalculated variables changed more than five percent. Of all valid cells (including zero) for microaggregated variables, less than one percent changed more than five percent while fewer than two percent of the recalculated variables did so.

For the ADSS Cost Study, means by facility type were compared pre- and post-microaggregation. Change in means by facility type ranged from -2.9 percent to +2.1 percent. Overall changes in means were negligible, which is the intended result of micro-aggregation.

Table 1. Overall effects of microaggregation and recalculation.

PHASE I FACILITY FILE		
	Microaggregated	**Recalculated**
Number of Variables	4	191
Record Count	2,394	2,394
Cells w/valid data (non-missing, non-0)	9,546	92,544
Cells w/missing data	0	289,062
Cells w/ data value=0	30	75,648
Total cells	9576	457,254
Change of > +/- 5%	82	3,304
Percentage (non-missing/non-0 cells)	0.859%	3.570%
Percentage (valid cells, including 0)	**0.856%**	**1.964%**
Percentage (total cells)	0.856%	0.723%

We further examined the impact to the data by comparing pre- and post-microaggregation ratios and means and by running a regression model on the pre- and post-microaggregated data to determine if significance results were comparable between the files.

Means were obtained by type of care and facility ownership for the microaggregated variables. The percent change in the means of these variables by both type of care and facility ownership ranged from zero to .9 percent, as shown in Tables 2 and 3. For the three total revenue variables that were impacted by microaggregation, the results are exactly the same for each variable. Therefore, only the result for one of these variables (D7) result is reported.

Table 2. Pre- and Post-Microaggregation Means By Type of Care.

PHASE I FACILITY FILE							
		Valid N		Mean		Absolute Difference	Percent Diff.
TYPCARE5 Type of care		Before	After	Before	After		
1 Hospital Inpatient Only	D7 Total subs abuse trt revenue	203	203	2658584.5	2680711.7	22127.3	0.8%
	B1j2 Total clients all care 10/1	203	203	18.4	18.4	0.0	-0.1%
2 Non - Hospital Residential Only	D7 Total subs abuse trt revenue	428	428	1176859.6	1169983.6	-6876.0	-0.6%
	B1j2 Total clients all care 10/1	428	428	43.8	43.8	0.0	0.0%
3 Outpatient Methadone Only	D7 Total subs abuse trt revenue	324	324	924848.3	924933.8	85.5	0.0%
	B1j2 Total clients all care 10/1	324	324	251.8	251.9	0.1	0.0%
4 Outpatient Non -Methadone Only	D7 Total subs abuse trt revenue	1083	1083	424329.1	424517.7	188.6	0.0%
	B1j2 Total clients all care 10/1	1083	1083	148.3	148.8	0.6	0.4%
5 Combination Facilities	D7 Total subs abuse trt revenue	356	356	1885023.6	1880021.3	-5002.3	-0.3%
	B1j2 Total clients all care 10/1	356	356	188.1	186.4	-1.8	-0.9%

Table 3. Pre- and Post-Microaggregation Means By Type of Facility Ownership.

PHASE I FACILITY FILE							
		Valid N		Mean		Absolute Difference	Percent Difference
A_6 A6. Type Of Ownership Of Facility		Before	After	Before	After		
1 Private For-Profit Organization	D7 Total subs abuse trt revenue	498	498	833230.4	838088.3	4858.0	0.6%
	B1j2 Total clients all care 10/1	498	498	145.2	146.4	1.3	0.9%
2 Private Non-Profit Organization	D7 Total subs abuse trt revenue	1478	1478	1040034.7	1037923.1	-2111.5	-0.2%
	B1j2 Total clients all care 10/1	1478	1478	127.8	128.4	0.6	0.5%
3 City / County Government Agency	D7 Total subs abuse trt revenue	249	249	1023422.0	1026405.5	2983.5	0.3%
	B1j2 Total clients all care 10/1	249	249	183.9	178.0	-5.9	-3.2%
4 State Government Agency	D7 Total subs abuse trt revenue	95	95	1349593.9	1355634.6	6040.8	0.4%
	B1j2 Total clients all care 10/1	95	95	103.1	103.1	0.0	0.0%
5 Federal Government Agency	D7 Total subs abuse trt revenue	63	63	2056990.0	2046533.0	-10457.0	-0.5%
	B1j2 Total clients all care 10/1	63	63	224.1	223.3	-0.8	-0.4%
6 Tribal Government	D7 Total subs abuse trt revenue	11	11	809306.2	813274.0	3967.8	0.5%
	B1j2 Total clients all care 10/1	11	11	68.2	67.9	-0.3	-0.4%

The *regression* model used the revenue variable "Other government funds" (D8G) as the dependent variable. This is a limited dependent variable in that roughly 86 percent of the 2394 programs in the sample database have an actual or implied zero (0) value for the amount of government funding. Therefore, an ordinary linear regression analysis of the full data is not appropriate and four regression analyses were tested. All analyses were done in STATA and incorporate the global sample weight variable (PH1FW0); however, the analysis did not include design effects for stratification. The data set was prepared with replicate weights for Balanced Repeated Replication analysis of complex sample design standard errors. This would require the use of Wesvar PC 4.0, which does not permit estimation of one of the models evaluated. Estimated coefficients computed in weighted analysis using STATA will exactly match those from the full analysis based on the complex sample design; however, the standard errors of the coefficients (shown in Table 6) are likely to be slight underestimates of the standard errors that would be obtained in an analysis that also included the stratification and weighting effects for the sampling of programs.

Model 1: Ordinary least squares regression on only the cases that have a nonzero amount for the government revenue variable. There are n=322 cases in this analysis.

Model 2: Ordinary least squares regression on only the cases that have a nonzero amount for the government revenue variable. The dependent variable is the natural log of the original non-zero government revenue amount. There are n=322 cases in this analysis.

Model 3: A Logistic regression model to analyze the probability that a program receives government revenue for its services. There are n=2394 cases in this analysis.

Model 4: A Tobit regression model for the left-censored (zero) dependent variable. There are n=2394 cases in this analysis.

Table 4 presents the results comparing the fit of each of these four models to the data before and after the microaggregation disclosure protection, showing that the regression model coefficients and the interpretation of the significance of the associated effects are quite robust against the microaggregation "blurring" of the data.

Table 4. Regression Model Test of ADSS Microaggregation.

	Model 1						Model 2					
	Ordinary Least Squares Regression[1] (D8G > 0)						Ordinary Least Squares Regression[1] of log(D8G) , (D8G > 0)					
	Before			After			Before			After		
Independent	Coeficient	Std. Err.	Sig.	Coeficient	Std. Err.	Sig.	Coeficient	Std. Err.	Sig.	Coeficient	Std. Err.	Sig.
b1a2	31439.38	5053.35	***	44674.75	5980.03	***	0.035	0.012	**	0.050	0.013	***
b1a2	3099.88	1736.91		3973.33	2092.24		0.019	0.004	***	0.017	0.005	***
b1h2	4025.55	785.49	***	3957.97	794.05	***	0.006	0.002	**	0.005	0.002	**
B1i2	613.88	333.15		804.65	375.48	*	0.001	0.001		0.003	0.001	***
a_4a	138645.11	92014.76		137918.6	114406.31		1.368	0.226	***	1.435	0.254	***
a_4b	-228279.2	114701.11	*	-212349.1	123279.31		-1.889	0.281	***	-1.834	0.273	***
a_4c	5409.22	55629.13		25567.77	100308.21		0.245	0.136		0.151	0.222	
a_61	-606374.9	1852.5.31	***	-565913.7	192445.71	***	-0.394	0.454		-0.425	0.426	
a_62	-653998.5	132995.11	***	-614072.1	14086.11	***	-1.017	0.325	**	-0.909	0.311	**
cons	792361.51	226134.81	***	671953.4	250034.61	***	11.591	0.555	***	11.512	0.554	***

Note[1]: (n = 322 cases)

	Model 3						Model 4					
	Logistic Regression[2] for Probability that D8G > 0 (recipiency)						Model 4: Tobit Regression[2] of D8G (left censored at 0)					
	Before			After			Before			After		
Independent	Coefficient	Std. Err.	Sig.	Coefficient	Std. Err.	Sig.	Coefficient	Std. Err.	Sig.	Coefficient	Std. Err.	Sig.
b1a2	0.032	0.004	***	0.026	0.004	***	22977.19	4808.56	***	24239.69	5458.15	***
b1d2	-0.001	0.001		0.001	0.001		-474.53	957398		-717.39	1249.77	
b1h2	-0.001	0.001		-0.001	0.001		-2677.04	656.78	***	-2871.59	711.08	***
b1i2	0.001	0.001	***	0.001	0.001	*	196.36	273.47		8.01	306.83	
a_4a	-0.277	0.045	***	-0.502	0.062	***	-196275	64173.82	**	-384277.9	99917.98	***
a_4b	-0.001	0.061		-0.062	0.067		160943.2	90743.59		172378.9	108183.21	
a_4c	0.125	0.051	*	0.353	0.057	***	-28987.88	61549.37		83193.49	90762.26	
a_61	-0.298	0.092	***	-0.286	0.093	***	-915518.5	152452.7	***	-874209	164155.81	***
a_62	-0.093	0.078		-0.104	0.079		-343161.6	104227.3	***	-330899.4	114515.41	***
cons	-1.567	0.127	***	-1.519	0.156	***	-925699.8	190982.8	***	-943137	245463.31	***

Note[2]: (n = 2394 cases)

*significant at the .05 level
**significant at the .01 level
***significant at the .001 level

Deletions

Any variables that could specifically identify a facility were removed from the file. These included variables such as facility name and address, facility director's name, name and address of parent organization, and National Master Facility Index (NMFI) identifiers. Also deleted were administrative variables such as interviewer initials and date and time of the interview and the "other, specify" variables that were provided as verbatim responses and had not been numerically coded. Client date of birth was also removed. One record was deleted from the Phase 1 facility file because it was either an extreme outlier or the revenue data had been coded or entered incorrectly.

Recodes

In addition to the variables that were recoded due to the microaggregation procedures, some variables were recoded to make them more analytically useful. For example, time intervals such as length of time for treatment, were recoded to a standard unit (e.g., a variable with responses of days, weeks, or months was recalculated to days). This was not possible for all time units because some variables had response options that could not be reduced to a standard unit such as *sessions*, days, weeks, etc. Also, records were randomized and facility and client identification numbers were removed and replace with sequential IDs, retaining the linkages between the files.

The codes for substance abuse and mental health disorders based on the Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria were recoded from the raw DSM codes into groups that made this variable more analytically useful. Table 7 shows the recoded diagnostic categories.

Table 7. Diagnosis recodes

ORIGINAL CODES	RECODES
0.00	0 No Diagnosis
291.00-291.99	1 Alcohol-induced Disorder
292.00-292.99	2 Substance-induced Disorder
303.00-303.89	3 Alcohol Intoxication
303.90-303.99	4 Alcohol Dependence
304.00-304.09	5 Opioid Dependence
304.20-304.29	6 Cocaine Dependence
304.30-304.39	7 Cannabis Dependence
304.10-304.19 304.40-304.99 305.10-305.19	8 Other Substance Dependence
305.00-305.09	9 Alcohol Abuse
305.20-305.29	10 Cannabis Abuse

ORIGINAL CODES	RECODES
305.30-305.49 305.70-305.99	11 Other Substance Abuse
305.50-305.59	12 Opioid Abuse
305.60-305.69	13 Cocaine Abuse
293.89 300.00-300.02 300.21-300.23 300.29-300.39 308.30-308.39 309.81	14 Anxiety Disorders
296.20-296.39 300.40-300.49 311.00-311.09	15 Depressive Disorders
293.81-293.82 295.00-295.99 297.10-297.19 298.80-298.89 297.30-297.39 298.90-298.99	16 Schizophrenia/Other Psychotic Disorders
296.00-296.09 296.40-296.79 296.80, 296.89 301.13	17 Bipolar Disorders
312.80-312.81 312.90-312.99 313.81 314.00-314.01 314.90-314.99	18 Attention Deficit/Disruptive Behavior Disorders
All other codes	19 Other Mental Health Condition
.01-289.99 320-997.99 V- and E-codes	20 Other Condition
Missing	-9 Missing

Excerpted from:

ALCOHOL AND DRUG SERVICES STUDY (ADSS)

DATA FILE USER'S MANUALS FOR THE

ADSS PHASE I, II, and III DATA FILES

Submitted to Brandeis University

by Westat

Under SAMHSA Prime Contract Number 283-92-8331

September-November 2000

Substance Abuse and Mental Health Services Administration

Office of Applied Studies

NOTE: The following information, specific to the Stratified Jackknife Factor and Finite Population Correction files, was excerpted from the codebooks for the ADSS Phase I-III facility and client data files.   For full documentation, please see the codebook corresponding to the ADSS data file to which these factor files will be applied.

WEIGHTING

[This is a summary only; please see codebooks for individual ADSS data files for full description of weighting].

Phase I

The Phase I sampling design incorporated a stratified random probability sample. Weights were developed for the Phase I sample to facilitate overall and by stratum estimates of facility- and client-level characteristics of the nation's substance abuse treatment system. Final Phase I weights were constructed in a multistep process involving calculation of initial base weights, trimming to guard against excessive influence by a few highly loaded facilities, adjustment for facility nonresponse, and poststratification adjustment to initial frame counts.

Because the Phase I sample was selected using a complex multistage design, resampling is the appropriate method of calculating the stability of computed statistics. Replicate weights based on the stratified jackknife procedure (JKn) are included in the ADSS Phase I dataset for the purpose of standard error calculations.

Phase II

Phase II weights, facility and abstract, were constructed for the entire Phase II sample based on type of care (residential, outpatient methadone, or outpatient nonmethadone), but without regard to Main Study/Incentive Study classification. Facility level weights are provided on the Phase II Administrator Interview File. Abstract level weights are provided on the Phase II main study abstract file and on the Phase II in-treatment methadone abstract file. The Phase II early dropout abstract file is not weighted.

Phase III

Phase III client weights were constructed for the Phase III Main Study interview file and for the Phase III In-Treatment Methadone Client file. The Phase III Incentive Study file and the Phase III outpatient nonmethadone early dropouts file are not weighted.

The major weighting steps were:

- Establishing client base weights,

- Raking, and

- Trimming.

REPLICATION THEORY

The basic idea behind replication is to select subsamples repeatedly from the whole sample, calculate the statistic of interest for each subsample, and then use the variability among these subsample or replicate statistics to estimate the variance of the full sample statistic. Different ways of creating subsamples from the full sample result in different replication methods. The subsamples are called replicates and the statistics calculated from these replicates are called replicate estimates. WesVar supports both balanced repeated and jackknife approaches.

The ADSS uses the general stratified jackknife (JKn) method. For a more detailed discussion of replication, its advantages and disadvantages, see Appendix A of the WesVar Complex Samples documentation.

The idea behind replication methods is to calculate the estimate of interest from the full sample, as well as from each subsample or replicate. The variation between the replicate estimates and the full sample estimate is then used to estimate the variance for the full sample. The variance estimator, $v(\hat{\theta})$, generally takes the form

$$v(\hat{\theta}) = c \sum_{g=1}^{G} f_g k_g \left(\hat{\theta}_{(g)} - \hat{\theta} \right)^2 \tag{5}$$

where

θ is an arbitrary parameter of interest

$\hat{\theta}$ is the estimate of θ based on the full sample

$\hat{\theta}_{(g)}$ is the g-th replicate estimate of θ based on the observations included in the g-th replicate

G is the total number of replicates formed

c is a constant that depends on the replication method (c=1 for Jkn method)

$v(\hat{\theta})$ is the estimated variance of $\hat{\theta}$

k_g are the JKn factors

f_g are the finite population correction factors.

The JKn factors are described below and are contained in the file JKN_FAC.DAT. For ADSS, there is a different file of JKn factors for each of Phase I, II, and III. The finite population correction (FPC) factors were created for each replicate for Phase I only. A file of the FPC factors is provided (FPC.DAT). For Phase I analyses, the FPC and JKn factors need to be attached within WesVar. The example that follows will show how the factors are attached. The effect of ignoring these factors is to overstate the variance.

Jackknife n (JKn)

The jackknife n (JKn) method can be used when the number of PSUs (VarUnits) in a stratum (VarStrat) is greater than or equal to 2. Therefore, the sample design for JKn is more general than for JK2 and BRR, which require exactly two VarUnits per stratum. The number of replicates, *G*, is equal to

$$\sum_{h=1}^{L} n_h$$

where *L* is the number of VarStrat and n_h is the number of VarUnits in stratum *h*. The maximum number of degrees of freedom is *G-L*. For ADSS Phase I, 200 replicates were created.

The general computations involved in forming the replicate weights in JKn were as follows. For the first replicate weight, the full sample of observations in the first VarStrat and VarUnit were multiplied by 0 and the weights associated with the other VarUnits in the same VarStrat were adjusted by $n_h/(n_h - 1)$ to account for reducing the sample. The weights of the observations in other VarStrat were not changed. The remaining *G*-1 replicates were formed in the same manner by systematically dropping each of the remaining VarUnits and computing the replicate weights as described for the first replicate.

The procedure generated JKn factors (k_g as shown in (equation 5)) that should be applied to the squared deviation of replicate *g* from the full sample estimate. The JKn factors are computed as $k_g = (n_{h'} - 1)/n_{h'}$, where *h'* identifies the stratum that is aligned with replicate *g*. Therefore, the factor for the *g*-th replicate weight depends on the number of unique values of VarUnit in VarStrat *g*.

About the Examples

This document contains examples that are intended to illustrate how to compute weighted estimates and standard errors for ADSS data using WesVar. The examples are from the Phase I facility-level data. They use the ADSS Phase I data from the SAS transport data set P1DELIV.XPT, JKn factors from the file JKN_FAC.DAT, and fpc factors from the file FPC.DAT. They illustrate how to create a WesVar data set from a SAS transport data set, the format in which ADSS files are delivered. Additionally, they show how to create a WesVar workbook to estimate totals and regression parameters and their associated variances, and then how to view the output from a workbook. Furthermore, the WesVar variances are compared to variances from SAS PROC MEANS.

IMPORTING THE SAS FILE

      Step 1. From WesVar's main screen, click the New WesVar Data File button or from the menu select **File ➤ New ➤ WesVar Data File.**

      Step 2. Select the file that you want to import and click **Open**. (Defaults for the import data file directory and for the WesVar data file directory can be specified in WesVar's Preferences) Choose the data set P1DELIV.XPT from the **Open** dialogue window. Browse for the folder that the file is in and be sure to change the "Files of type:" to either *.xpt (transport files) or *.* (all files). Any SAS for Windows files (.sd2) must be converted to .ssd or Transport files (.xpt) before being imported. Converting to a .ssd file can be done in SAS using the libname statement: libname *libref* v604 <*'SAS-data-library'*>. Converting to a .xpt file can be done using the libname statement: libname *libref2* xport <*'SAS-data-library'*>; along with the PROC COPY procedure (PROC COPY in=*libref1* out=*libref2*; select P1DELIV; run;).

      Figure 4-1 shows the WesVar Data File screen displays.
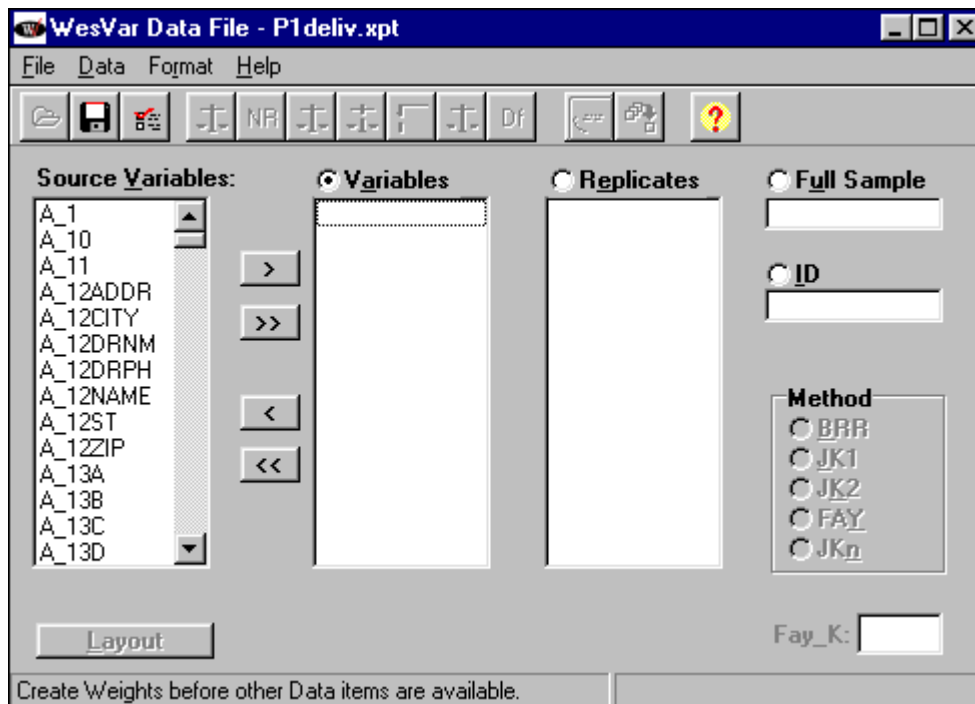
Figure 4-1. WesVar Data File Screen

      On this screen you can identify variables, replicate weights, the full sample weight, ID variables, and the replication. An ID variable is one that is used solely to identify the case or record. If you have an ID variable and designate it as such on the WesVar Data File screen, it cannot be used in any Table or Regression request. The ID variables are retained on the WesVar data file and can be extracted later.

The left-hand column lists the source variables that were on the imported file.

Step 3. Click the appropriate box to identify variables, replicate weights, the full sample weight, or ID variables.

Step 4. Move variables from the Source Variables list to the appropriate box by double-clicking the variable, using the arrow buttons, or dragging.

As you move the variables, they will disappear from the left-hand column and appear in the appropriate box. It may be easiest to move the ID, Full Sample, and Replicate weights first, and then move the remaining variables simultaneously to the Variables box using the double arrow button.

You do not have to move all of the source variables into the WesVar data file, but any variables left in the Source Variables list cannot be added to the WesVar data file after it is created.

Step 5. For ADSS data, choose the JKn replication method by clicking on **JKn** in the Method box.

Step 6. When all variables have been selected and moved, save the imported file as a WesVar file. From the menu select **File ➤ Save**. The Save As dialog box displays.

Step 7. To save the file, either click the **Save As** icon on the toolbar or select **File ➤ Save** from the menu. If you are saving the file for the first time, the Save As dialog box appears. Keep the default file name "P1DELIV" or type in a new name for the file. WesVar will convert the file from an SAS transport *.xpt file format to a WesVar *.var file format.

The WesVar Data File screen in Figure 4-2 shows the variables that were identified and the new file name in the title bar on the screen.
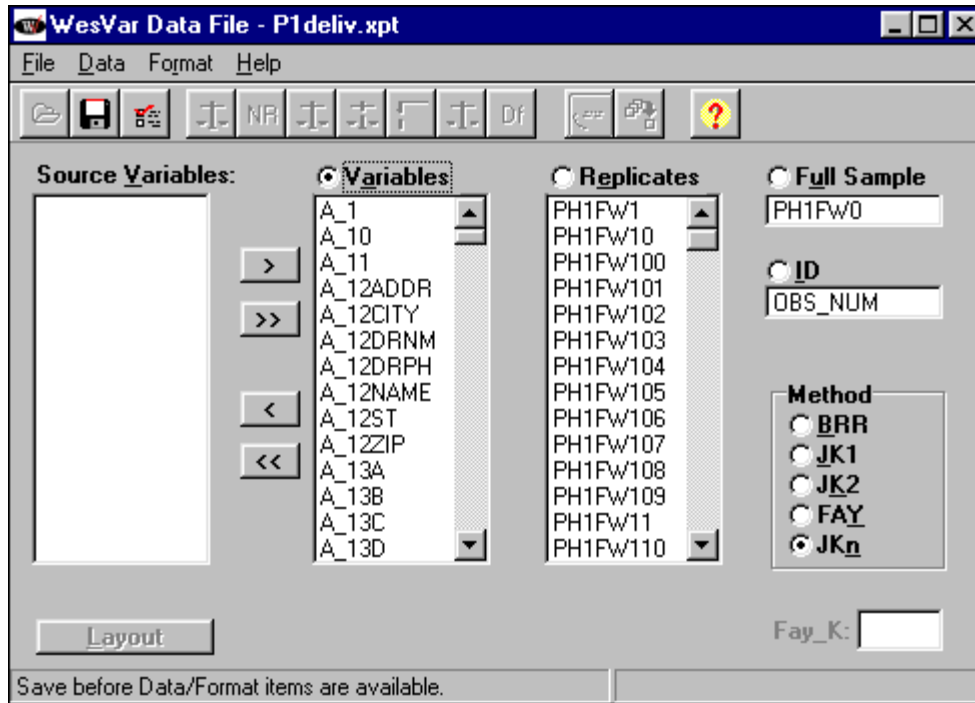
Figure 4-2. WesVar Data File with Replicates

Attach Factors

The Attach Factors feature is an advanced way to attach fpc factors and JKn factors.

To attach factors:

Step 1. Open a WesVar data file and from the menu select Data ➤ Attach Factors.

Step 2. Open the external file that contains the fpc factors. Highlight the column for fpc factors, click Open, and select the file fpc.dat. Repeat analogously to import JKn factors from the file JKN_FAC.DAT. The first factor in the file will be linked to the first replicate, the second factor to the second replicate, etc.

After these factors are imported, the screen will look like Figure 4-3.
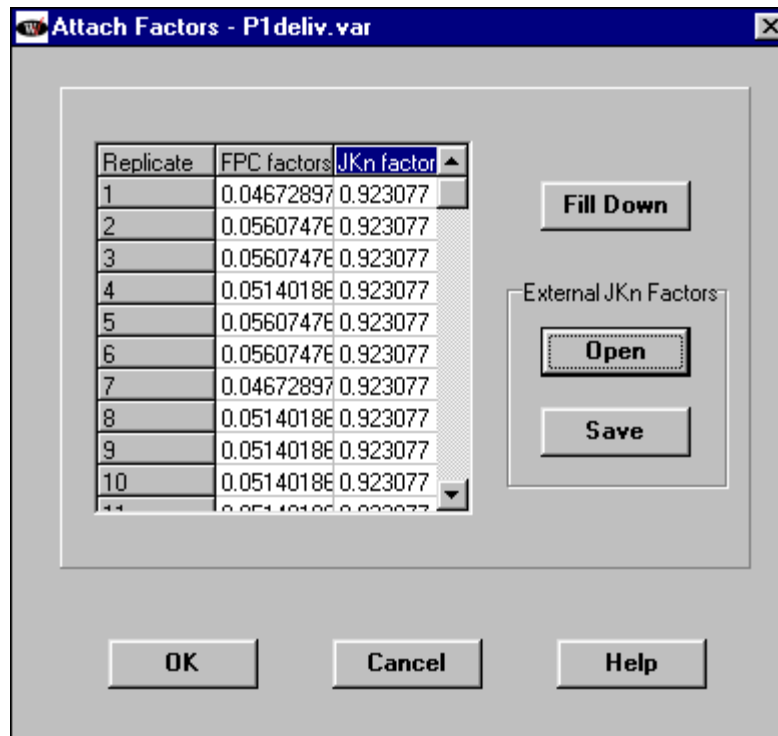
Figure 4-3. Attaching Factors

Step 3. When all factors have been set, click **OK** then **Save**.

Your WesVar data file has now been created. Exit from the data file screen by double-clicking on the WesVar icon in the top left corner, or by selecting **File ➤ Close**. To use this .var file, click on **New WesVar Workbook** or select **File ➤ New ➤ WesVar Workbook**. Find the .var file you have created and click **Open**.

Creating a Table

Click on **Table** on the right side of the screen. Edit the table request by clicking on it and changing the name on the right side of the screen, if you wish. By clicking on **Generated Statistics** and **Output Control**, you may specify options for this table request. For global changes, type **Ctrl-P**. To create a frequency of a discrete variable, highlight **Tables** on the left side of the screen, search for and double-click on the variable of interest under **Source Variables** on the right side (see Figure 4-4). It will then become selected. Click on **Add as New Entry** to incorporate the table request.

Suppose you want to estimate the total number of facilities and the total number of clients by treatment type (*TYPCARE5*). Since the total number of facilities is estimated by the sum of weights (seen in

Figure 4-4), select the **Value** box under **Sum of Weights**. For population estimates of the number of clients, use *B1J2* (Total Clients all Care) and select the **Value** box under **Analysis Variables**.

In addition to population totals, WesVar allows the option of returning percentages—overall, row, and column. This is done by checking the appropriate dialog boxes on the right side of the screen of Figure 4-4.
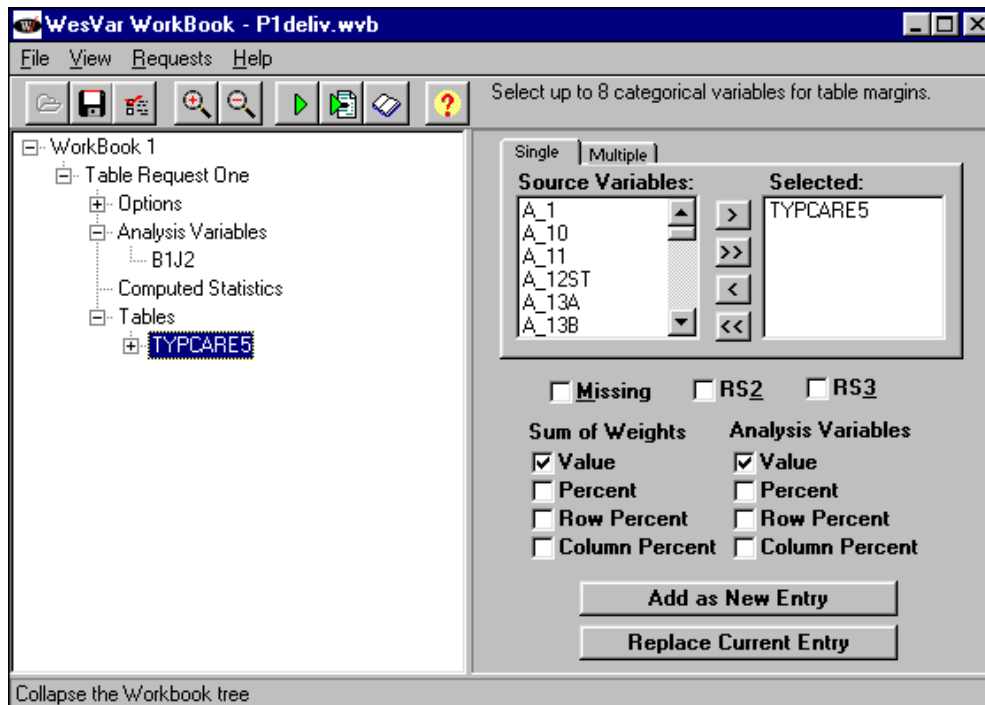

Figure 4-4. Example of Table Request

Viewing the Output

Run the request using the green triangle button on the menu bar. When WesVar has completed the table, the button (an open book) for viewing the table turns from gray to white. Click on the open book button to view the output. Expand the tree on the left side of the output screen and click on *TYPCARE5* (Facility Type of Care). The table appears on the right side of the screen (see Figure 4-5). Errors, if any, appear as a red exclamation point next to the name of the table, and a message at the bottom right explains the problem.

The output gives estimates of the number of facilities by type of care and total number of clients by facility type of care. Marginal values are also given to estimate the entire population.

Other values such as standard error and sample size can be reported, but they must be specified under the **Generated Statistics Option** of the Table Request.

Figure 4-5. Viewing the Table Output

Creating a Regression

Suppose you want to create a regression to model the relationship between number of full-time staff (*A_9I1*) and total admissions (*C2F1*). To create a regression in WesVar, simply click on **Regression** at the workbook node.

Under Models, select *A_9I1* as the dependent and *C2F1* as the independent variable from the list of source variables provided and click on **Add as New Entry** to incorporate the selection into the Regression Request (see Figure 4-6).
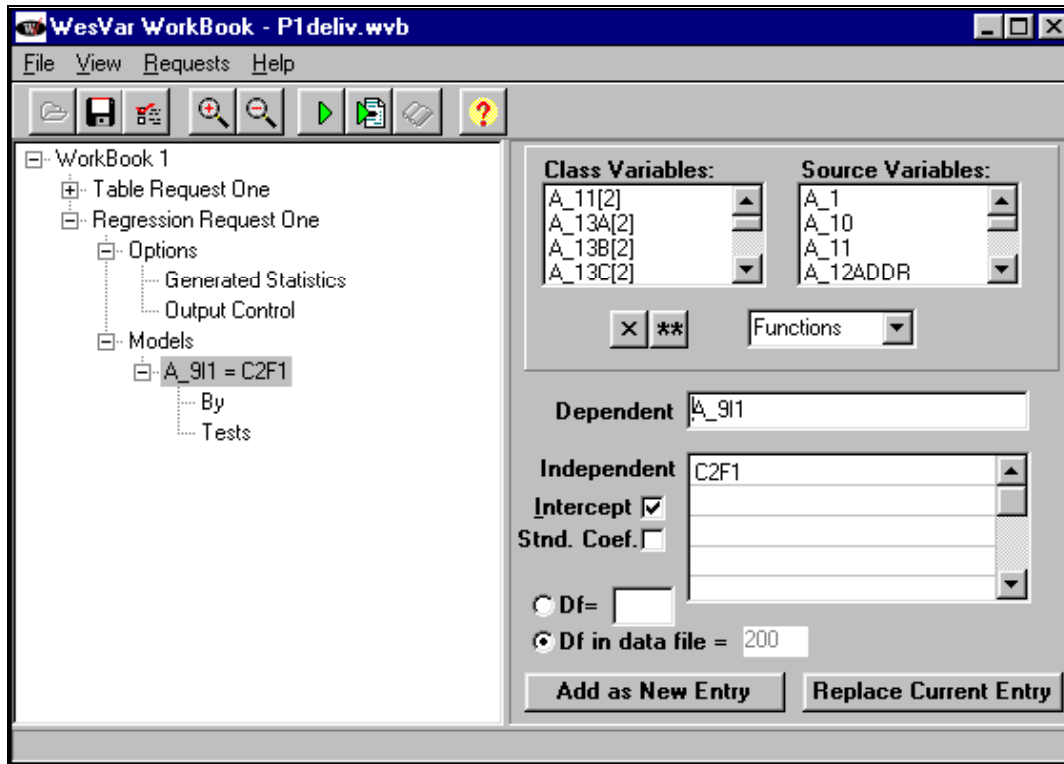
Figure 4-6. Incorporating the Regression Request

Run and view the regression output in the same way as in the table example. Expand the menu on the left side and highlight **Estimated Coefficients** (see Figure 4-7). The regression output is typical, reporting estimates, standard errors, test statistics, p-values, and an R^2 value.

Figure 4-7. Viewing the Regression Output

Highlight the **File** menu for printing and exporting the newly created table.

COMPARING WESVAR TO SAS

It is often desirable to compare the standard error given by WesVar (taking the complex sample design into account) to a simple random sample standard error. Table 4-1 compares standard errors of mean number of clients (*B1J2*) by facility type (*TYPCARE5*). The SAS standard errors were found using PROC MEANS with the options VARDEF=WEIGHT and STD, the CLASS statement *TYPCARE5*, and *PH1FW0* as the weight. The resulting standard deviations were then divided by $\sqrt{n}$ to produce the numbers in Table 4-1.

Table 4-1.   Comparison of standard errors produced by SAS and WesVar for the levels of Typecare

TYPECARE	SAS	WesVar
1	1.4743	0.99152
2	1.6798	1.35701
3	8.9465	8.53646
4	4.0425	3.66633
5	13.3445	11.80387
marginal	3.1347	2.98808

The difference between the standard errors from SAS and WesVar shows the effect that the ADSS Phase I sampling and weighting procedures have on the variances.

ANALYSIS ISSUES

Degrees of Freedom

The default degrees of freedom for WesVar tabular and regression analysis is the total number of replicates.[1] This may be assumed for large domains such as ADSS analytic strata, since the number of active replicates at each stratum level is relatively large. However, for small domains, the approximate degrees of freedom need to be specified. The degrees of freedom can be specified in the **Options** panel for tables and the **Models** panel for regression. To count the number of active replicates, produce a frequency on the variance unit and variance stratum variables for the subset. The number of active replicates equals the number of unique combinations of variance strata and variance units remaining on the data. The approximate degrees of freedom may be calculated by subtracting the number of unique variance strata remaining on the data. Recall that the variance stratum and variance unit together identify the replicate to which the sample unit is aligned.

Accounting for the Imputation Error Variance in ADSS Data

In Phase I of ADSS, imputation was performed on a select group of variables that had item nonresponse. Treating imputed values as if they were actually observed or reported may lead to a significant understatement of the variance of the estimate. Several methods have been developed to account for the effects of imputation error in variance estimation. The All-Cases Imputation (ACI) method was developed

---

[1] The default degrees of freedom for tabular requests may be modified by the user on the **Tables(2)** tab under **File…Preferences**. The options are Infinite, Number of Replicates, and User Specified.

recently (Montaquila and Jernigan, 1997) and we chose to apply the method to ADSS (Krenzke, Mohadjer, Montaquila, 1998).

Since the missingness rates in the clients, staffing, and much of the admissions items was small, it was assumed that the imputation error variance was negligible for these items. For selected items with higher nonresponse rates (total revenues, total costs, and other key items within the admissions, revenue, and cost blocks), the imputation error variance was estimated using the ACI method. This method consists of imputing for all cases, not just the nonrespondents, and using the imputed and observed values for the respondents to estimate the imputation error among the nonrespondents.

The model-assisted ACI approach assumes ignorable nonresponse and provides an unbiased estimate of the variance of the mean under generalized conditions. In general, the ACI estimator of the total variance has three components. The first component is the sampling error variance (S^2), the second is the imputation error variance (I^2), and the third is the imputation error covariance. The third component is considered negligible for ADSS since the donors could be used only once. Therefore, the ACI estimator of the variance of the mean reduces to two terms:

$$\hat{v}_{ACI}(\bar{y}_{st}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left[\frac{(1-f_h)}{n_h}\hat{v}\left(y_{hi}^*\right) + \frac{m_h}{n_h^2}\hat{v}\left(\tau_{hi}\right)\right];$$

where, $\hat{v}(y_{hi}^*)$ = the sample variance among the actual and imputed values of the characteristic y in stratum h, and $\hat{v}(\tau_{hi})$ = the sample variance among the respondent imputation errors in stratum h. Both variance terms were computed using WesVar. The stratified jackknife technique was used to compute the variance components.

To simplify the computation of variances in the presence of imputation error, the approach recommended for ADSS is to incorporate the imputation error variance by using a variance inflation factor (*VIF*). This factor can be multiplied by the variance (computed by treating imputed values as if they were observed) after the calculation of the jackknife variances in WesVar.

Total Variance = $V(\hat{\theta}) * VIF$, where $V(\hat{\theta})$ is the resulting variance from WesVar and VIF is shown in Table 4-2.

Table 4-2 shows the *VIFs* for each of the ADSS Phase I items that contain imputed values. While some of the *VIFs* were computed using the ACI method, other *VIFs*, as noted, were generalized through simply using the ratio of the *VIF* to the imputation rate, for a closely related item. The amount of

missingness filled in by statistical imputation methods is shown, as well as the variance inflation factor (*VIF), VIF = (S^2 + I^2)/S^2*. For financial items, most facilities were able to report revenues and costs relating to substance abuse only; however, others could not. Therefore, before using statistical methods for imputation, the financial item values were transformed into values representing substance abuse only, using *D10BOX* and *D10PC* for revenues, and *D18BOX* and *D18PC* for costs.

Alternative Software for Analyzing Survey Data

This section summarizes two alternative software packages, SUDAAN and Stata, that were developed for analyzing data from complex surveys. These were already referenced in the Introduction. Both packages can be used with ADSS data.

Table 4-2.   Variance Inflation Factors (VIFs) for ADSS Phase I items that contain imputed values

Item	VIF
Items not relating to type of care	
C4A	1.06*
C4ANUM	1.04
C4B	1.11
C4BNUM	1.07
D7	1.02
D14	1.01
D15A	1.03
D15B	1.03**
D15C	1.03
D16D	1.08
Items relating to hospital inpatient treatment	
D12A	1.07
D16A	1.25
Items relating to residential treatment	
D12B	1.01
D16B	1.01
Items relating to outpatient treatment	
D12C	1.03
D16C	1.00
Items not relating to type of care	
D13A	1.03
D16C2	1.00
Items relating to methadone treatment	
D13B	1.03
D16C1	1.00

* The *VIF*s were not directly computed for *C4A* and *C4ANUM*. The ratio of the *VIF* to statistical imputation rate for the corresponding *C4B* terms was used to approximate the *VIF* for the *C4A* items.

**The *VIF*s were not directly computed for *D15B* and *D15B*. The ratio of the *VIF* to statistical imputation rate for *D15A* was used to approximate the *VIF* for *D15B* and *D15C*.

SUDAAN

The section is intended to help readers who are already somewhat familiar with SUDAAN in their use of SUDAAN when analyzing ADSS Phase I data. SUDAAN requires the selection of a DESIGN option and the identification of variables in a number of required and optional command statements, such as the NEST command. The section describes the possible choices that are appropriate with ADSS data and indicates some of the strengths and weaknesses associated with them.

Choice of Design

In SUDAAN**,** three DESIGN options may seem appropriate for use with ADSS Phase I data, one taking a replication approach and the two others making use of the Taylor's series expansion method. These three options are discussed below:

DESIGN = JACKKNIFE

This option does not allow the current replicate weights on the file to be read in. Using DESIGN = JACKKNIFE (replication) is a reasonable option, but it should be used cautiously since the approach of replicating final full sample weights may cause serious overestimates of sampling error. Recent work by Brick, Morganstein, and Barrett (1999) has shown some serious overestimates of variance estimates for totals, and to a lesser extent for means and proportions, for three national surveys using this technique. Results depend on the correlation of the survey items with the weighting variables, levels of nonresponse, and effects of poststratification. A possible correction would be to re-poststratify the resulting replicate weights. However, since one would not be able to read back into SUDAAN the re-poststratified replicate weights, DESIGN = JACKKNIFE may not be an appealing option. The use of variables *TOTCNT* and *SMPCNT* is omitted for the JACKKNIFE option (refer to the paragraph 'Population and Sample Size Variables' in this section for the definition of *TOTCNT* and *SMPCNT*). Therefore, the option JACKKNIFE will produce overestimates of variance where the sampling fraction is high in noncertainty strata. One can use the NEST command to give levels of the design (stratum and primary sampling unit). A brief description of the use of the NEST command is provided in the paragraph 'The Nest Command' in this section. For DESIGN = JACKKNIFE, one can use the ADSS variables *VARSTRAT* and *VARUNIT*, which were used as stratum and PSU variables for producing stratified jackknife replicates for use in WesVar.

DESIGN = UNEQWOR

Another option is DESIGN=UNEQWOR, which uses Taylor's expansion for estimating variances. This option, however, may not be practical since the computation of joint probabilities under systematic sampling is very complex for analysts to incorporate. In order to incorporate the finite population correction (FPC) factor, two ADSS variables may be used in the DESIGN = UNEQWOR option. For the UNEQWOR option, *TOTCNT* is required and *SMPCNT* is an option (refer to the paragraph 'Population and Sample Size Variables' in this section for the definition of *TOTCNT* and *SMPCNT*). One can use the NEST

statement to give levels of the design (stratum and primary sampling unit). A brief description of the use of the NEST command is provided in the paragraph 'The Nest Command' in this section.

DESIGN = WR

The most reasonable SUDAAN option to use is DESIGN = WR (Taylor's expansion) if the certainty facilities are placed in their own strata, even though the sample was selected without replacement. The option WR will produce overestimates of variance where the sampling fraction is high in noncertainty strata, since the use of variables *TOTCNT* and *SMPCNT* for accounting for the finite population correction factors is unavailable. Refer to the paragraph 'Population and Sample Size Variables' in this section for the definition of *TOTCNT* and *SMPCNT*. One can use the NEST statement to give levels of the design (stratum and primary sampling unit).

The Nest Command

To analyze ADSS Phase I data, the required NEST command can specify *VARSTRAT* and *VARUNIT* as the variables designating stratum and PSU, respectively. For a further explanation of these variables and the methodology used to create them, see Krenzke and Mohadjer (2000). For the design options UNEQWOR and WR, one can redefine the stratum variable by placing the certainties (identified by the second digit of *VARSTRAT*=1) in their own strata, and use *OBS_NUM* as the primary sampling unit, which is the facility.

Population and Sample Size Variables

The frame totals, *TOTCNT*, which is a value on each record on the data file, represent the number of facilities on the sampling frame (including ineligibles) for the sampling stratum corresponding to the facility data record. Similarly, the sample counts, *SMPCNT*, represent the number of unit respondents, for the sampling stratum corresponding to the data record.

To account for the effects of poststratification, when developing the replicate weights for ADSS using the stratified jackknife, poststratification was done first by adjusting the weights of the entire ADSS Phase I sample to the frame totals, which also include ineligible facilities. In SUDAAN, for DESIGN = WR, one may use *TOTCNT* along with the statement POSTVAR to capture the effects of poststratification. This was investigated by Flores-Cervantes, Brick, and DiGaetano (1999) for the 1997 National Survey of

America's Families (NSAF) for the Urban Institute. SUDAAN's poststratification option brought the variance estimates back in line with WesVar estimates, so they concluded that the estimates from SUDAAN were reasonable.

Note that the resulting point estimates do not change when using the POSTVAR option; only the variance estimates change. In fact, all three software packages should generally produce the same point estimates. When using the poststratification option in SUDAAN, it may be necessary to redefine the strata and the frame counts if the certainty facilities are grouped in their own strata. If certainties are grouped in their own strata, then to arrive at frame totals for the noncertainty strata, the number of certainties in the corresponding sampling stratum should be subtracted from the corresponding frame counts, *TOTCNT*. For the certainty strata, the frame counts should be set to 1. Poststratifying to frame counts that include ineligibles in SUDAAN (for variance purposes only) may also lower the variance estimate beyond the effect it has on the WesVar variances. The aforementioned result is due to SUDAAN treating the poststratification totals as the truth when, in fact, there is uncertainty surrounding the count of eligible facilities on the frame. This result supports the recommendation of using WesVar to compute variance estimates when using ADSS data.

Stata

In the Stata software, the Taylor's expansion methods are used to estimate variances. The software offers several *svy* statements to cover several different types of analyses, including means, totals, and ratios. The stratum population sizes are needed if the FPC factors are to be incorporated. The function *svyset* sets up the sampling strata and the PSU identifiers. Since poststratification may have a significant effect on the variance, and since Stata does not incorporate such an effect into the variance estimates, results from Stata should be interpreted cautiously. Flores-Cervantes, et al. (1999) also mention that Stata does not have the poststratification option, so it was not as useful for their purposes. In addition, as in SUDAAN, the variance estimates do not reflect the effects of nonresponse weighting adjustments and weight trimming. Variance estimates are generally higher than those from WesVar and SUDAAN (if the POSTVAR option is used as an option).

Comparing WesVar, SUDAAN, and Stata

Resulting variances are different depending on the software package being used. The magnitude of the differences between the results from the software packages depends on several factors, including type of analysis, impact of systematic sampling, and impact of weighting procedures. It is important for the user

to explain how the standard errors were computed. Furthermore, data users are encouraged to consult the software developers of WesVar, SUDAAN, and Stata. As noted at the start of this chapter, WesVar is the recommended choice for analyzing ADSS data since the sample and replication scheme were designed with WesVar in mind.

Broene and Rust (1998) prepared a Westat report to the National Center for Education Statistics (NCES) documenting their evaluation of statistical software packages for NCES data sets. At the time of the evaluation, both SUDAAN and Stata used a linearization approach to variance estimation; SUDAAN's latest version includes replication methods. Broene and Rust's paper mentions that SUDAAN is probably the most powerful of the three packages, but may be the most difficult to learn. They conclude that WesVarPC (soon to be WesVar 4.0) was both easy to learn and powerful but lacks some of the model fitting capabilities that SUDAAN has. Furthermore, they mention that Stata is more limited in its survey data analysis capabilities and can be slower to run, but it does enable one to easily plot and examine predicted values and residuals when model-fitting. They mention that all three packages compute standard errors for proportions and for continuous statistics such as means, totals, ratios, and differences in these quantities. For categorical analysis, SUDAAN and WesVar were recommended.

Since the time of the Broene and Rust report, several enhancements were made to each software package. Table 4-3 compares some current features of each package (WesVar 4.0 (due for release in the second half of 2000), SUDAAN 7.5, and Stata 6.0). Note that Stata is fully programmable, meaning that, if Stata does not already have a specific function, a program may be created to satisfy individual needs.

Table 4-3.   Analysis capabilities for WesVar, SUDAAN, and Stata

	WesVar	SUDAAN	Stata
Standard errors and design effects for means, totals, proportions, ratios	X	X	X
Standard errors for Quantiles	X	X	X

Finite population correction factor:
    1st stage only, equal probabilities of selection
    1st stage only, unequal probabilities of selection
	X	X
		X
			X

Linear regression	X	X	X

Logistic regression:
    Dichotomous
    Polychotomous
	X
	X
		X
		X
			X
			X

Probit models			X
Loglinear models		X	X
Tests of independence in tables	X	X	X
Linear contrasts, differences	X	X	X
Survival analysis		X	X
Graphics			X
Batch processing available	X	X	X
Output useful for importing into spreadsheets	X	X	X
Estimates and confidence Intervals for odds ratios in logistic regression	X	X	X
Tests in logistic regression models	X		X
Adjust replicate weights for nonresponse	X		
Correlation matrices (in addition to covariance matrices)	X		X
Design effects	X	X	X

APPENDIX A - REFERENCES

Batten, H.L., et al. (1992). *Drug Services Research Study: Final Report Phase II,* U.S. Department of Health and Human Service.

Batten, H.L., et al. (1993). *Drug Services Research Study: Final Report Phase I (Revised),* U.S. Department of Health and Human Service.

Brick, J.M. and Morganstein, D.R. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the Bureau of the Census 1996 Annual Research Conference.*

Brick, J.M., Morganstein, D.R., and Barrett, B. (1999, November). Variance estimation: Estimating the effects of weight adjustments. Presented at the Washington Statistical Society Seminar.

Broene, P. and Rust, K. (1998). Strengths and limitations of using SUDAAN, STATA, and WesVarPC for computing variances from National Center for Education Statistics data sets. *Westat report prepared under contract for NCES.*

Butler, M.A. and Beale, C.L. (1993). Rural-Urban Continuum for Metro-Nonmetro Counties, U.S. Department of Agriculture: Washington, DC.

Flores-Cervantes, I., Brick, J.M. and DiGaetano, R. (1999). 1997 NSAF variance estimation. Report No. 4, *NSAF Methodology Reports: Assessing the New Federalism*, The Urban Institute, Washington, DC.

Kalton and Kish, L. (1984). Some efficient random imputation methods. Communications in Statistics, 13(16), pp. 1919-1939.

Krenzke, T., and Mohadjer, L. (2000). Sample Design and Weighting Report for Phase II and III of ADSS prepared for SAMHSA: Westat

Krenzke, T., Mohadjer, L., and Montaquila, J. (1998). Generalizing the imputation error variance in the Alcohol and Drug Services Study. To appear in the *American Statistical Association's 1998 Proceedings of the Section on Biometrics*.

Mohadjer, L., Yansaneh, I., Krenzke, T. and Dohrmann, S. (2000). Sample design, selection and estimation for Phase I of ADSS. (Prepared under contract to the Substance Abuse and Mental Health Services Administration). Rockville, MD: Westat

Montaquila, J. and Jernigan, R. (1997). Variance estimation in the presence of imputed data. *American Statistical Association's 1997 Proceedings of the Section on Survey Research Methods,* pp. 273-278.

Montaquila, J. and Ponikowski, C. (1995). An evaluation of alternative imputation methods. American Statistical Association's 1997 Proceedings of the Section on Survey Research Methods, pp. 281-286.

Substance Abuse and Mental Health Services Administration. (1992). *Drug Services Research Survey: Final Report: Phase I (Revised) and Phase II,* February, 1992.

Westat (1998). WesVar Complex Samples User's Guide. Chicago: SPSS, Inc.