

APPENDIX TO USER'S GUIDE

Some survey analysts may wish to correct standard errors for weighting and clustering using statistical packages such as SUDAAN or Stata. This appendix proposes one method to prepare the data for such an analysis.

We have included the following variables on each data file (continuing/new caregiver, separated caregiver, and focal child):

SCRID: Screener ID

PU: Primary frame unit

SEGID: Segment identification number

SITE: City

1=Boston

2=Chicago

3=San Antonio

STR: Race/ethnicity stratum

B=Non-Hispanic Black/African-American

W=Non-Hispanic White

H=Hispanic/Latino

The firm conducting the survey, Research Triangle Institute (RTI), constructed eight sets of block groups from all of the blocks in the three cities in the 1990 Census. Each set ranked all the block groups in a city in descending order of the poverty rate of children in a particular race-ethnic group. In Boston, three such sets were compiled and ranked --one each for Non-Hispanic Whites, Non-Hispanic Blacks, and Hispanics. Three analogous sets were compiled for Chicago, and two sets for San Antonio--Non-Hispanic Blacks and Hispanics. Only block groups falling below a specified poverty level were retained in the sampling frame. The variable SITE refers to the city from which the block groups were drawn. The variable STR refers to the racial/ethnic composition of the block group.

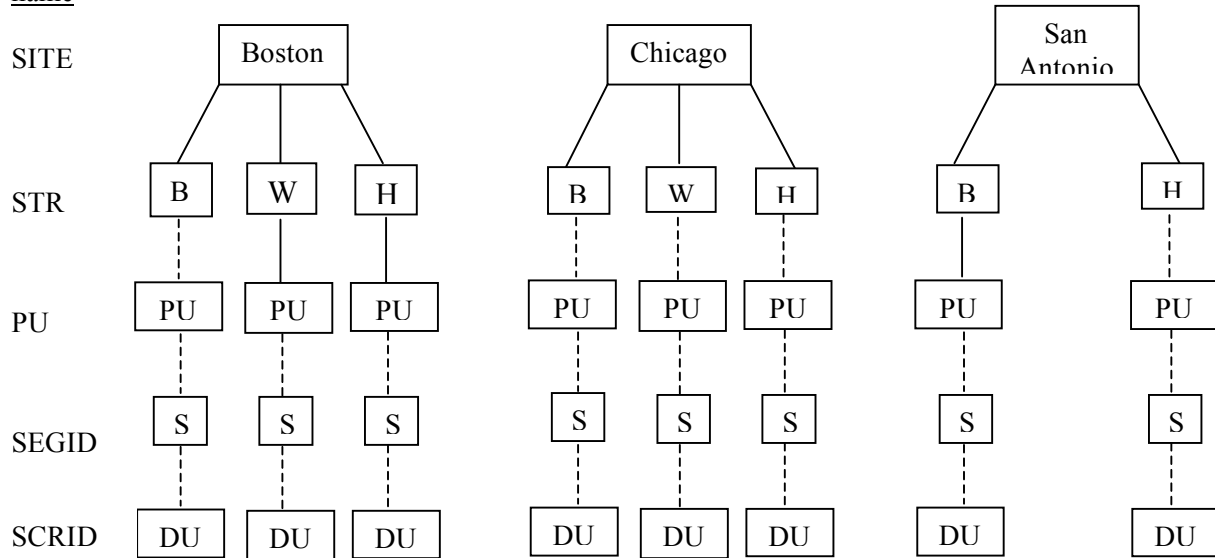
The block groups in the sampling frame are referred to as primary frame units and are represented in the variable PU. In five of the eight sets described above, a random set of the PUs was selected with probability proportional to size. These five sets are referred to as "non-certainty strata," meaning that not all of the eligible PUs were selected into the sample. In the other three sets, all of the PUs were selected. These are certainty strata. Within these strata, all values of the variable PU are blank.

From each stratum, the selected PUs were divided into segments, which are areas of a size typically regarded as convenient for surveying and generally consist of 90-120 dwelling units. A set of segments was chosen randomly from the selected PUs. All selected segments were then counted and listed (i.e., interviewers visited the segments, counted the housing units, and wrote down the addresses of all occupied dwelling units). Segments are represented in the variable SEGID. A random sample of dwelling units (identified by street addresses) was then selected

from within each segment. These dwelling units are the households that appear in our sample. The unique identifier SCRID represents each household.

The schematic below shows how households were selected into the sample. The solid lines represent points at which all eligible units were included. The dashed lines represent points where only a subset of units (whether PU's, segments, or dwelling units) was selected.

Variable
name



The method proposed below to account for clustering was developed by the survey contractor. Different methods are used for households in certainty and non-certainty strata. Data users may wish to use other techniques with which they are familiar. For a general discussion of data preparation for complex survey data analysis, see Eltinge and Sribney (1996), Chapter 16 in Levy and Lemeshow (1999), and StataCorp (2003).

For certainty strata (White and Hispanic strata in Boston (site=1, str= "W" or "H") and the Black stratum in San Antonio (site=3, str= "B")):

Sort all households by SITE, STR, SEGID, SCRID. This sorts the data into a geographically ordered list, with households ordered by segment within a site/race-ethnic group. The variable PU takes no value for these cases. Go down the list and form pairs of dwelling units in a new variable called STRATA. Number these pairs from 1 to N/2. Then call each of the dwelling units within a stratum a CLUSTER. The clusters will take on the values of 1 or 2.

For example, assume you are working with a sample that includes 1000 cases from the certainty strata. The first two observations will have the value 1 on the new variable STRATA, the next two observations will have the value 2, and so on. The last two cases will have the value 500. Within each pair, the first observation will have the value 1 on the new variable CLUSTER. The

second observation will have the value 2. A list of these data would have the following appearance:

SCRID	SITE	STR	PU	SEGID	STRATA	CLUSTER
1201010S	1	H	.	1201	1	1
1201020S	1	H	.	1201	1	2
1201030S	1	H	.	1201	2	1
1201040S	1	H	.	1201	2	2
1202010S	1	H	.	1202	3	1
1202020S	1	H	.	1202	3	2
.						
.						
.						
2561010S	3	B	.	2561	500	1
2561020S	3	B	.	2561	500	2

No pair should cross between two site/race-ethnic groups. If there is an odd number of dwelling units in an area, the last dwelling unit should be assigned a cluster value of 2 and attached to the last pair in its site/race-ethnic group, so that the last “pair” would actually include three observations.

For non-certainty strata:

Sort all households by SITE, STR, SEGID, and PU. Here, definitions of clusters and strata are not based on the dwelling units and dwelling unit pairs within the non-certainty strata. Rather, PUs and PU pairs are used for cluster and strata definition. Again, the pairing of PUs should be done within a common area, so that pairs do not cross area type.

For example, assume you are working with a sample that includes 1000 observations from the non-certainty strata. Among those 1000 observations, 200 PUs are represented. The number of PU pairs that would emerge from this sample would be $(\# \text{ of PUs})/2$, or $200/2=100$. All of the observations within the first PU in a given pair would carry a value of 1 on the variable CLUSTER. The observations within the second PU would carry a value of 2. A list of these data would have the following appearance:

SCRID	SITE	STR	PU	SEGID	STRATA	CLUSTER
1401010S	1	B	1	1401	1	1
1401020S	1	B	1	1401	1	1
1401030S	1	B	1	1401	1	1
1401040S	1	B	1	1401	1	1
1402010S	1	B	5	1402	1	2
1402020S	1	B	5	1402	1	2
1403010S	1	B	7	1403	2	1
1403020S	1	B	7	1403	2	1
1403030S	1	B	7	1403	2	1
1404010S	1	B	9	1404	2	1

1404020S	1	B	9	1404	2	2
1404030S	1	B	9	1404	2	2
.						
.						
.						
2261010S	3	H	140	2261	100	2
2261020S	3	H	140	2261	100	2

In Stata's svyset command, the analyst may set the variable STRATA as the strata identifier variable, and the variable CLUSTER as the PSU (cluster) identifier variable.

References

Eltinge, J.L. and W.M. Sibney (1996). svy3: Describing survey data: sampling design and missing data. *Stata Technical Bulletin* 31: 23-26. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 235-239.

Levy, Paul S. and Stanley Lemeshow (1999). *Sampling of Populations: Methods and Applications*. 3rd ed. New York: John Wiley & Sons, Inc.

StataCorp (2003). *Stata Survey Data Reference Manual, Release 8*. College Station, TX: Stata Press.