

Introduction to SUDAAN® for survey data

This course and other resources

This is a course on how to begin using the statistical programming package SUDAAN, specifically the SAS callable release 9.0.0. This course assumes no previous experience with SUDAAN, although students should

- Know basic SAS syntax since we will be using SAS callable SUDAAN in class (Students that need review can attend one or more Introductory SAS classes), and
- Intend to analyze data collected via surveys.

Note: David Chae has some tips that have proven to be extremely useful. Whenever you see this symbol ? it signifies a DCT (David Chae tip)!

SUDAAN is a single program comprising a family of analytic procedures. SUDAAN procedures are used to analyze data from complex sample surveys and other observational and experimental studies involving repeated measures and cluster-correlated data. Here we exclusively present examples of analyses from survey data.

Useful resources:

? On-line help from the SUDAAN web site <http://www.rti.org/sudaan/onlinehelp/FlashHelp/SUDAAN.htm>

? Article about analysis of survey data and the use of statistical software Brogan, Donna J. **Pitfalls of using standardized statistical software packages for sample survey data**, *Encyclopedia of Biostatistics*, Peter Armitage and Theodore Colton, eds., section “Design of Experiments and Sample Surveys”, Paul Levy, ed., John-Wiley 1998.

Topics in this class include

- Introduction to cluster-correlated data in the survey setting
- Inputting data
- Descriptive procedures like CROSSTAB and DESCRIPT
- Brief theory of parameter and variance estimation using SUDAAN
- Modeling procedures like REGRESS, and RLOGIT
- Use of survey design variables

What’s different about survey data?

Most standard statistical methods assume you are conducting a **simple random sample** (SRS) to collect the data to be analyzed. This requires that the whole population is accessible in the sampling frame and then a sample is chosen randomly.

Introduction to SUDAAN for survey data

Example: In the Mega Millions 12-state lottery you choose 5 numbers out of a possible 56. The whole frame is represented by the 56 numbers possible. If you let the machine pick the five numbers for you, it supposedly results in a SRS.

The SRS assumption does not typically hold for the collection of survey data.

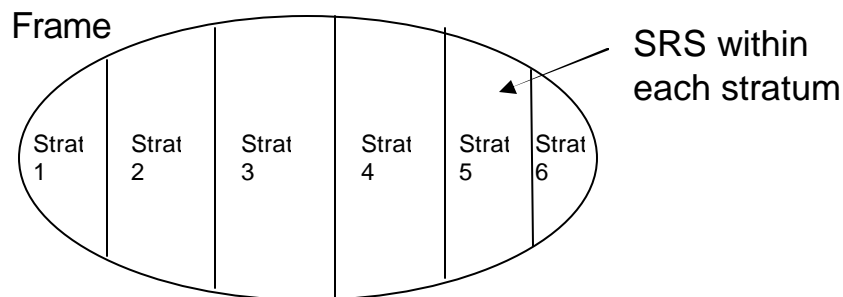
Brogan (1998) wrote:

Standard statistical software packages generally do not take into account four common characteristics of sample survey data:

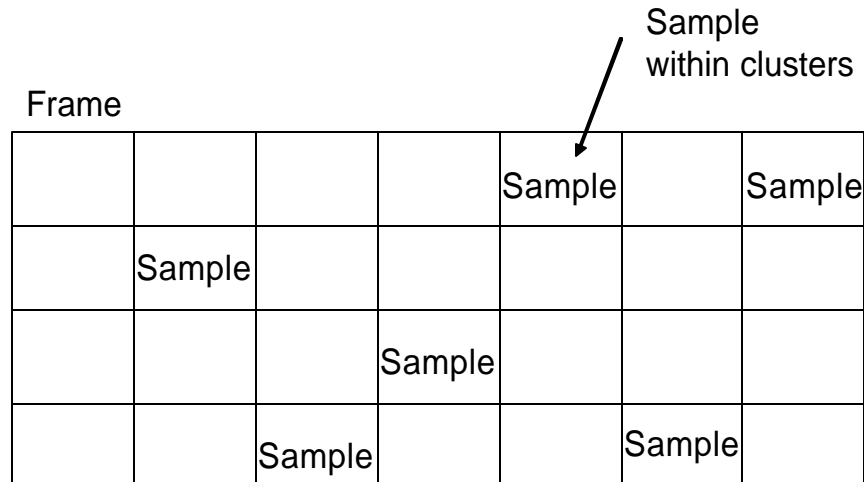
- (1) unequal probability of selection of observations,
- (2) clustering of observations,
- (3) stratification, and
- (4) nonresponse and other adjustments.

As part of the survey data collection, samples could have one or more of the following features.

1. **Stratified sampling:** The sampling frame is partitioned into groups or *strata*, and the sampling can be performed separately within each stratum (e.g. stratification by Asian ethnic identity to assess specific characteristics of those subpopulations.)



2. **Cluster sampling:** The sampling frame is composed of units that are difficult to enumerate, so the sampling frame is constructed to identify groups or *clusters* of enumeration units without listing explicitly the individual enumeration units (e.g. it is difficult to enumerate all individual people in the USA, so clusters are constructed.)



3. **Sampling weights:** A sampling *weight* is an adjustment to the individual unit such that when the unit measurement is multiplied by the weight and an overall estimate is computed, that estimate will represent the entire frame. (e.g. If it is known that the entire USA is 51% female and 49% male, but the actual sample drawn for analysis is 60% female and 40% male, weight male responses more heavily to compensate for under-representation in the sample.)

These features of survey design require us to apply appropriate methods that account for stratified sampling, cluster sampling, sampling weights, or any combination of these.

NOTE: Although most statistical packages can perform weighted analysis, if the other sampling characteristics are not accounted for appropriately, it could lead to “biased point estimates, inappropriate standard errors and confidence intervals, and misleading tests of significance” (Brogan, 1998).

Which statistical software packages can account appropriately for survey design characteristics?

- SUDAAN
- Stata
- Possibly others?

Analysis with Categorical Variables

A categorical variable has discrete categories associated with it.

Examples:

Diseased / Not Diseased at a specified time point

Variable **DIS** is defined as

1 = Diseased

0 = Not Diseased

Note: This is an “indicator” or “dummy variable” because it takes on values 0 or 1, exclusively.

Economic status

Variable **ESTAT** is defined as

1 = Below Poverty Level

2 = Poverty Level

3 = Working Income Level

4 = Middle Income Level

5 = Upper Income Level

These levels are internally defined.

Access SUDAAN via remote terminal server

Exercise 1:

Open a window to the NLAAS remote terminal server. Open SAS. In the Enhanced Editor, type the following code.

```
* Set-up environment;  
libname myfolder 'D:\temp';
```

Check that your dataset SAMPLE is there by navigating through the Explorer Window.

```
* First, look at the contents of the SAMPLE dataset;  
proc contents data=myfolder.sample;  
run;
```

Prevalence

SUDAAN’s CROSSTAB procedure produces weighted frequency and percentage distributions for one-way (single-variable) and multi-way (multi-variable) tabulations.

Exercise 2: Compare SAS’s PROC FREQ with SUDAAN’s PROC CROSSTAB with a one-way distribution for two variables, MARSTAT and WORKSTATUS.

Type and then run the following code.

```
* SAS: PROC FREQ;
```

Introduction to SUDAAN for survey data

```
proc freq data=myfolder.sample;
  tables marstat workstatus;
title EX 2 FREQ;
run;

* SUDAAN: PROC CROSSTAB;
proc crosstab data=myfolder.sample filetype=sas design=srs;
  class marstat workstatus;
  print / style=nchs;
title EX 2 CROSSTAB;
run;
```

Question: In this sample, how many people have employment status "Out of labor force"?

Since this is a one-way distribution, the PROC CROSSTAB results for row, column and total percentages are equal to each other.

Let's go over parts of the PROC CROSSTAB syntax.

<code>data=</code> <i>datafile-name</i>	What data is to be use for this procedure?
<code>filetype=</code> <i>filetype</i>	What type of file is the data-file? SAS
<code>design=</code> <i>design-spec</i>	What design is this? SRS (Simple random sample)
<code>class</code> <i>var(s)</i>	What variables would we like to compute frequencies on? MARSTAT and WORKSTATUS
<code>print / style=</code> <i>style-type</i>	Print options are optional - try omitting this and see the difference. ?NCHS is a more compact style.
<code>title</code> <i>title-text</i>	Title can be part of any procedure to help identify output

Exercise 3: Add survey design characteristics to PROC CROSSTAB.

```
* First sort data by design variables;
* ?Note: The dataset name must be 8 characters or less;
proc sort data=myfolder.sample out=sample;
  by strata cluster;
run;

* Add a NEST and WEIGHT statement;
* Note the change in the DESIGN specification from SRS to WR - see
explanation in notes;
proc crosstab data=sample filetype=sas design=wr;
  nest strata cluster;
  weight wgtvar;
  class marstat workstatus;
  print / style=nchs;
title EX 3;
run;
```

Question: What is the estimated proportion of the population

Introduction to SUDAAN for survey data

(prevalence) that are married and what are the confidence limits associated with the estimate?

Look at additional PROC CROSSTAB syntax.

<code>nest strata var clustervar</code>	What are the strata and cluster design characteristics? (Note: sort by these ahead of time.)
<code>weight weight-var</code>	What is the weight variable?
<code>design=design-spec</code>	What design is this? WR (default - with replacement)

Exercise 4: Create a two-way table.

```
* Exercise 4: Create a two-way table;
* ?Note: MARSTAT is a "dummy variable" taking values (0,1) so
  change that to a (1,2) variable to include in the TABLES statement;
data sample2;
  set sample;
  marstat = marstat + 1;
  label marstat = 'Marital status: 1=not married 2=married';
run;
```

```
* Exercise 4: Create a two-way table;
proc crosstab data=sample2 filetype=sas design=wr;
  nest strata cluster;
  weight wgtvar;
  format best;
  subgroup marstat workstatus;
  levels 2 3;
  tables marstat * workstatus;
  print / style=nchs;
title EX 4;
run;
```

?NOTE: SUDAAN must have categorical variables in a 1,2,3,... format for many of the statements to work.

Question: What is the total estimated percent of the population that are employed?

Question: Among married people, what is the estimated percent of the population that are unemployed?

Look at additional PROC CROSSTAB syntax.

<code>format best</code>	The format statement specifies how statistics will be displayed in the output. ? Format best is useful if some of the estimates are small.
<code>subgroup var(s)</code>	What categorical variables will be used? (Note:

Introduction to SUDAAN for survey data

	must be used with TABLES statement)
levels var-levels	How many categorical levels correspond to the subgroup variables?
tables var * var	What two (or more) variables do you want crossed for frequency computations? (Must have a SUBGROUP and LEVELS counterpart)

What do you do when you want to subset the data?

Exercise 5: Repeat the Exercise 4 CROSSTAB , but subset the data on the portion of the sample that is not obese (OBESE = 0).

```
* Exercise 5: Create a two-way table subset on OBESE=0;
proc crosstab data=sample2 filetype=sas design=wr;
  nest strata cluster;
  weight wgtvar;
  subpopn obese=0;
  subgroup marstat workstatus;
  levels 2 3;
  tables marstat * workstatus;
  print / style=nchs;
title EX 5;
run;
```

Question: How many people in the sample are members of the non-obese (OBESE=0) subpopulation?

Look at additional PROC CROSSTAB syntax.

subpopn *logical-expression* SUDAAN only uses the records for which the logical-expression is true. ?Note: Using SUBPOPN statement is not equivalent to using a subset of the data file where the observations you wish to exclude have been deleted. Differences will be evident in estimates of standard errors; using SUBPOPN corresponds to the assumption that even if there are no individuals in a cluster in the sample, there may be some in the universe, and an appropriate contribution to the estimated variance must be calculated.

? SUBPOPN can be used when there are missing values in the variables of interest. One way to use it is to create an indicator variable (e.g. values = 1) for whenever *all* values for the collection of variables to be used for the analysis are *not missing* and an alternative value for all other cases (values = 0 or values = .). Then use the indicator variable (SUBPOPN values = 1) to ensure that standard errors will be estimated appropriately even with missing observations.

Introduction to SUDAAN for survey data

Add a test for association.

Exercise 6: Does employment status differ significantly by marital status using a chi-square test?

```
* Exercise 6: Add a chi-square test for association;
proc crosstab data=sample2 filetype=sas design=wr;
  nest strata cluster;
  weight wgtvar;
  subpopn obese=0;
  subgroup marstat workstatus;
  levels 2 3;
  tables marstat * workstatus;
  test chisq;
  print / style=nchs;
title EX 6;
run;
```

Question: What is the p-value for the chi-square test of association?

Look at additional PROC CROSSTAB syntax.

```
test test-name
```

Use the TEST statement with the appropriate test-name to cause CROSSTAB to print test statistics.

Logistic Regression

Developing a logistic regression model requires a two-category outcome (dependent variable). For SUDAAN, that outcome variable must take on the values (0,1). (Note: the idiosyncrasy that independent variables must have categorical values 1,2,3, ... still holds. Weird!)

Exercise 7: Is age, marital status, or work status associated with a person being over weight? Use a logistic regression model.

```
* Logistic Regression;
* Exercise 7: Dependent variable=OVERWGT, Independent variables=AGE
MARSTAT WORKSTATUS;
proc rlogist data=sample2 filetype=sas design=wr;
  nest strata cluster;
  weight wgtvar;
  class marstat workstatus;
  refllevel marstat=1 workstatus=3;
  model overwgt = age marstat workstatus;
  print / style=nchs;
title EX 7;
run;
```

Question: What is the estimated beta coefficient for being married?

Introduction to SUDAAN for survey data

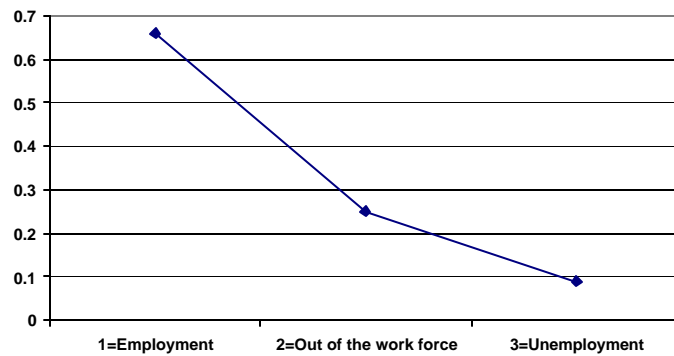
See additional RLOGIST syntax. Note: this is not LOGISTIC but **RLOGIST**. If you use PROC LOGISTIC, you will be calling a SAS procedure, not a SUDAAN procedure.

```
class var(s)           CLASS identifies categorical independent
                        variables
reflevel logical-      REFLEVEL sets reference categories for all
expression             the variables listed in CLASS
model dep-var = ind-var(s) MODEL creates the model statement
```

Interaction

Could there be an interaction between marital status and working status with regard to being over weight? What's an interaction? Here's an example with fictitious data.

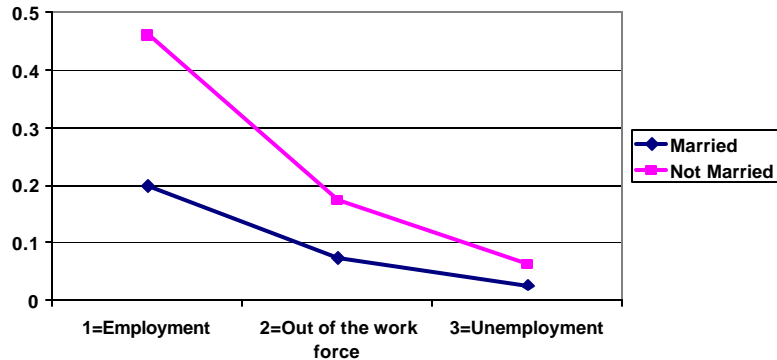
Proportion Over Weight by Employment Status



**Proportion Over Weight by Employment Status and Marital Status
No Interaction**



Proportion Over Weight by Employment Status and Marital Status Interaction



The last plot of an interaction between marital status and employment status with regard to being over weight shows a compounded effect. It could be said that the effect of employment status is **moderated by** marital status.

Enter interaction terms by crossing (*) them in the model.

Exercise 8: Add the interaction term MARSTAT * WORKSTATUS into the previous model.

```
* Exercise 8: Add an interaction term MARSTAT * WORKSTATUS;
proc rlogist data=sample2 filetype=sas design=wr;
  nest strata cluster;
  weight wgtvar;
  class marstat workstatus;
  refllevel marstat=1 workstatus=3;
  model overwgt = age marstat workstatus marstat*workstatus;
  print / style=nchs;
title EX 8;
run;
```

Question: What is the estimated odds ratio for being overweight when a person is married and employed?

Analysis with Continuous Variables

A continuous variable is one that takes on many values over a range. Typical examples include age, income, systolic blood pressure, height, and scales.

Exercise 9: What are the mean and standard error of the mean for the variables AGE, DISTRESS, and EDUCATION?

```
* Exercise 9: Compute descriptive statistics for continuous variables
AGE DISTRESS EDUCATION;
```

Introduction to SUDAAN for survey data

```
proc descript data=sample2 filetype=sas design=wr;
  nest strata cluster;
  weight wgtvar;
  var age distress education;
  print / style=nchs;
title EX 9;
run;
```

Question: What is the mean and 95% confidence bounds for the scale of distress?

Additional statements for the DESCRIPT procedure.

```
var var(s)                                List the variables for which mean and
                                           standard errors are to be computed.
```

Perform a T-TEST or ANOVA by using a linear regression model using the REGRESS procedure. ? Note: you can also use the DESCRIPT procedure for T-Tests.

Exercise 10: Is the mean of distress significantly different for married versus unmarried people? $H_0: \mu_M = \mu_U$ versus $H_a: \mu_M \neq \mu_U$

```
* Exercise 10: T-Test for mean distress scale by marital status groups;
proc regress data=sample2 filetype=sas design=wr;
  nest strata cluster;
  weight wgtvar;
  class marstat;
  relevel marstat=1;
  model distress = marstat;
  print / style=nchs;
title EX 10;
run;
```

Question: What is the p-value associated with this t-test?

Note that MARSTAT is the group variable by which we will compare distress means. It is categorical, while the variable DISTRESS is continuous.

Now consider a multivariate linear regression model.

Exercise 11: Are marital status, age, and the assessment of the quality of one's relationship with his/her children linearly associated with level of distress?

```
* Exercise 11: Multivariate linear regression - dependent=DISTRESS,
independent=MARSTAT AGE RELATE;
proc regress data=sample2 filetype=sas design=wr;
  nest strata cluster;
  weight wgtvar;
  class marstat workstatus;
  relevel marstat=1 workstatus=3;
  model distress = marstat age relate;
  print / style=nchs;
```

Introduction to SUDAAN for survey data

```
title EX 11;  
run;
```

Question: In the results of your model, how would you interpret the beta coefficient for AGE?