Using SAS™ to Implement Jackknife and Bootstrap Methods
For Survey Data

Tony Ye, Senior Statistician
Environmental Risk Analysis, Inc.

Ruby Lin, Biostatistician
Baxter Healthcare Corp.

Abstract

Analyzing data collected from surveys usually requires using special statistical packages (such as SUDAAN™). Although one can use SAS™ to generate point estimates, the variance associated with the estimate is usually overestimated because of the complicated survey structure. Development of bootstrap and jackknife theory provides an empirical solution for analyzing survey data. Fast growth of computing power made it feasible to use these techniques. We developed a SAS™ Macro to analyze National Health Interview Survey (NHIS) and compared our results to the one obtained by using SUDAAN™. For simple means, this SAS macro replicates SUDAAN™ results. For more sophisticated statistical modeling, this SAS™ Macro enables users to fit a larger range of models than using SUDAAN™ or using any other survey analysis package. In this article, we will also discuss how to use this macro to implement various statistical models to other national surveys.

Introduction

National surveys, such as National Health Interview Survey (NHIS), provide a rich set of information for health care research. For example, the annual national health interview survey covers topics from health care utilization to health habits. These data are useful not only to identify potential health risk factors but also to generate disease prevalence representing the entire U.S. population.

National surveys usually adopt complex survey design. A typical design is multistage sample design. In National Health Interview Survey, the United States is considered as 2000 Primary Sampling Units. A PSU consists of a county or a group of contiguous counties. These PSUs are clustered into 62 strata. 4 PSUs are selected from each stratum. We will discuss the survey structure in details in the next section.

Due to the complex survey structure, it is required to use special statistical software package to calculate the variance of a statistics generated using the survey data. Among the statistical package dealing with survey data, SUDAAN™ is one of the most popular products, which offers a large range of modeling options. SAS™ currently does not provide special procedure to correctly handle survey data.

The difference of variance estimation between SAS™ and SUDAAN™ is demonstrated in the following chart. In this example, we calculated average body weight for U.S. male age 50-65. Both SAS™ and SUDAAN™ gave the same average (192.5 pounds). However, the standard error estimated from SUDAAN was 0.74, while standard error estimated from SAS was 0.61. In this simple example, SAS™ under-estimated the standard error by nearly 20%.

Figure 1

To a SAS user, there are two limitations of using special survey analysis software such as SUDAAN. The first is that you may need to transfer your data back and forth from SAS to the survey analysis software. The second is that SUDAAN provides a limited number of statistical procedures. You may need to fit a statistical model that SUDAAN does not support. In such case, you can not rely upon either SUDAAN or SAS™ to provide the reliable statistical inference.

A closer look of the Survey Structure

Before we go into the alternative solution, the jackknife approach, we need to take a closer look of the survey structure to get a better understanding of where the extra variance comes from. The following chart demonstrates a typical national survey structure.

Survey Structure of National Health Interview Survey

As shown in Figure 2, in the 1994 survey, the survey conductor partitioned the entire U.S. into 62 strata. For example, Northern California may be stratum #1, and Southern California may be stratum #43. Within each stratum, there are about 30 primary sample units (or PSUs). A PSU consists of a county or a group of contiguous counties. In the 1994 NHIS, the agency selected 4 PSUs within each of the 62 strata. Within each PSU, the agency then selected more than 100 household as survey sample.
Intuitively we can see why the SAS™ would generate a different standard error than SUDAAN™ under the above survey structure. Let's assume the agency increase the number of sampled household units within each PSU from 100 to 10000. The SAS™ would generate a smaller standard error with a factor of 1/10. However, for a statistics calculated using this survey, most of the variance results from using only 4 PSU to generate estimate for each stratum. No matter how many households we select within each PSU, we can not reduce the uncertainty of using the 4 PSUs to estimate for 30 PSUs in each stratum.

The Solution: Jackknife

With the understanding of the survey structure, we are ready to move on to the alternative approach—the jackknife approach. Jackknife and bootstrap are statistical methods used to generate statistical inference. These methods do not require analytical representation for the variance. Instead, these methods may need large amount of calculation. Thanks to the fast growing computing capability, computing intensity is no longer a barrier to execute such calculation. To correct use the jackknife or bootstrap method requires correctly identifying the source of variance. In our case, it is to correctly generate jackknife sample using the same structure as the original national survey.

There have been more than 20 journal articles discussing the use of jackknife and bootstrap methods for survey data. I listed 4 of them in the reference. In this article, we used a jackknife approach. The approach includes four steps.

Step 1: Generate A Jackknife Sample
Step 2: Calculate Statistics Using a Proper Statistical Model and Keep the Results
Step 3: Repeat Step 1 and 2 Until Exhausting All Jackknife Samples
Step 4: Calculate the Variance of the Statistics Using the Kept Results from Each Run.

In the following section, using the example of calculating the average body weight for male 50-65, I will go into details discussing each of the four steps.

Step 1: Generate A Jackknife Sample

As shown in the Figure 2, the 1994 NHIS survey contains 62 strata, and each stratum contains 4 PSU. Totally we have 248 PSUs. In each jackknife sample, we delete one PSU. Totally we have 248 jackknife samples. Unlike normal sample, the survey sample requires one additional step to generate a jackknife sample. The following chart helps us to explain this step.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Design of the Jackknife Algorithm</th>
<th>National Health Interview Survey, 1994</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iteration</td>
<td>Strata</td>
<td>PSU</td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>-----</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>X</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

This additional step is to inflate the statistical weight for the rest of PSUs that belong to the same stratum of the deletion. For example, when we generate the first jackknife sample, we delete the subjects from contained in (stratum 1, PSU1). The first stratum is now under-represented because of having only 3 PSUs. We then inflate statistical weight for the rest subject in stratum 1 by a factor of 4/3 to overcome the under-representation due to deleting the first PSU. By the same time, we keep the statistical weight unchanged for subjects in other strata.

Step 2 & 3: Calculate Statistics Using a Proper Statistical Model and Keep the Results

In this step, you can fit any statistical model you think is proper for the data. You can use logistic regression, proportional hazard model etc. In our case, we calculate the simple mean of body weight using the jackknife samples. Compared to any survey data analysis software, Jackknife provides more options for selecting statistical models.

Step 4: Calculate the Variance of the Statistics Using the Kept Results from Each Run

Assuming that the statistics generated using each jackknife sample are $X_1, X_2,...,X_{248}$, the standard error generated using the jackknife methods can be expressed as the following formula.

$$\text{std.err} = \sqrt{\frac{3}{4} \sum_{i=1}^{248} (x_i - \bar{x})^2}$$

There are two features of this formula worth mentioned here. The first one is that the variance of the statistics is calculated as the sum of the squared differences. This sum should not be divided by 247. Intuitively this is because each jackknife sample differs from the original sample only by one PSU. One would anticipate the statistics generated using each of the jackknife samples to be very similar to the statistics generated using the original sample.

The second one is the factor $\frac{3}{4}$. This factor arises from the fact we delete one PSU (therefore we left 3 PSUs in the same stratum) to generate a jackknife sample. The proof of this formula is out of the range of this article. Readers can find the proof in the reference articles.
SAS™ Macro to Implement the Jackknife Program

We wrote a SAS™ Macro to implement this Jackknife algorithm. The first section of the program is to generate variables representing the sample survey structure. The second section is a SAS™ Macro that fits the statistical model on each of the jackknife sample. We simply used PROC MEANS in our analysis. User can change the PROC MEANS with any modeling procedure. The third section is to calculate the variance and confidence interval of the statistics. We attached the sample code at the end of this article.

Conclusion

In our example, we have 6,775 male subjects age 50-65. The jackknife algorithm took about 7 minutes for a Pentium 233 desktop to execute. The standard error estimated using the jackknife algorithm is 0.74, replicating what was generated by SUDAAN™.

Given the current computation advances, it is feasible to use the bootstrap/jackknife methods to generate statistical inference. Using these methods not only reduces the data management effort of transferring between different software systems, but also provides more modeling options. The jackknife/bootstrap method may be the only way to obtain a confidence interval for statistics estimated using a complicated statistical model.

Reference


5. Current Estimates From the National Health Interview Survey. 1994 Vital and Health Statistics Series 10 No. 193

Appendix

Sample Program for Jackknife Algorithm

/* Read in NHIS 1994 Data */
filename asc "\nhhis1994\boot\test.exp*;
DATA IN.test;
INFILE asc; INPUT
/* Source file: PERSONSEX.DAT */
/* Records: all */
@1 SEX $CHAR1.
@2 AGE $CHAR2.
@4 WEIGHT $CHAR3.
@7 STRATUM $CHAR3.
@10 strate $CHAR2.
@12 cpsu $CHAR1.
@13 NFTA $CHAR3.
*/

run;
data analysis:
set in.test;
if $0<age<65 and sex=1;
bodywt = weight +0 ;
t=wtfa+0;
strate=strate +0;
cpsu=cpsu +0;
run;

/* Generate Variables Needed for the jackknife */
proc sort data=analysis out=anal2 nodupkey;
by strata cpsu;
run;
data design:
set anal2;
by strata cpsu ;
retain psucount 1 ;
retain rep 0 ;
if first.strata then psucount=1;
else psucount=psucount+1;
rep=rep+1;
run;

proc means data=design nway noprint;
class strata ;
var psucount; output out=design2 max=tpsuc;
run;
data design3;
merge design design2;
by strata ;
factor=(tpsu-1)/tpsuc;
keep strate cpsu rep factor tpsu;
run;
data analysis;
merge analysis (in=anal2) design3 (keep=strate cpsu rep factor tpsu);
by strata cpsu;
if in1=1;
run;

/* Jackknife Macro */
%macro jack:
proc delete data=anal2; run;
&do i = 1 to 248;

data design4;
set design3;
if rep = &i; keep strata;
run;
data jack1;
merge analysis design4 (in=in2) ;
by strata ;
if rep= &i then newwt=0;
else if in2=1 then newwt=wt * tpsu/(tpsuc-1) ;
else newwt=wt;
one=1;
run;

/* Here We Used Proc Means, You can Change to any Proc You Think is Proper */
proc means data=jack1;
weight newwt;
var bodywt one;
output out=out1 mean=bodywt x sumy totalwt;
run;
proc append base=out2 data=out1;
run;
&end ;
%mend;
%jack;

/* Calculate Standard Error and Confidence Interval */
data out2:
set out2;
one=1;
run;
proc means data=out2 noprint;
var bodywt one;
output out=out1 mean=mean1 one;
run;

data result;
merge out2 out1:
by one;
diff = bodywt - mean1;
diff2 = diff * diff * 3/4;
run;

proc means data=result noprint;
var diff2 mean1;
output out=present sum=varr x mean=y mean;
run;

data present;
set present;
stderr = sqrt(varr);
L95 = mean - stderr*1.96;
H95 = mean + stderr*1.96;
run;

Title "Average Body Weight, For Male Age 50-55";
Title2 "WHO 1994";
proc print data=present;
var mean stderr L95 H95;
run;