Paper 266-27

How Complex Can Complex Survey Analysis Be with SAS®?

J.M. Gossett, P. Simpson, J.G. Parker and W.L. Simon
University of Arkansas for Medical Sciences, Little Rock, AR

ABSTRACT

Introduction: In version 7.0, SAS® included routines for analyzing complex survey data with the SURVEYMEANS® and SURVEYREG® procedures. These additions are welcome since SAS is the program of choice for many analysts and avoiding purchasing and learning other programs is preferable. However the SAS routines, which use Taylor linearization for estimating variances, are limited in their capabilities. There are several survey data sets, particularly government public release data, which include replicate weights. For these studies it is possible to extend the standard procedures using data step programming to calculate better variance estimates.
Aim: To show that SAS can be used for complex survey data sets that have replication weights, particularly for balanced repeated replication (BRR) and jackknife type II (JK2). Specifically to show that it is possible to use SAS to calculate properly weighted variance estimates for summary statistics such as the mean, and median using JK2 estimates.
Method: We will demonstrate that fairly simple algorithms can be implemented to estimate variances using replicate methods. The results will be shown to be equivalent to those obtained from SUDAAN® or WesVar®.
Conclusion: It can be seen that this approach allows a wider use of SAS in complex surveys.

INTRODUCTION

In version 7.0, SAS included routines for analyzing complex survey data with the SURVEYMEANS and SURVEYREG procedures. These additions are welcome since SAS is the program of choice for many analysts and it is desirable to avoid purchasing and learning specialized survey software programs such as SUDAAN®, WesVar®, or PC Carp®. However the SAS routines, which use Taylor linearization *(TL)* for estimating variances, are limited in their capabilities. TL obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975).

There are many complex survey designs that are not covered by the SAS/SURVEYMEANS procedure. The assumption that first-stage sampling is "with replacement" is not true in general. However, for many large scale surveys, the estimates are approximately correct. Indeed, a design might have strata with clusters or PSUs within strata, and samples without replacement taken from the PSU. If the variance estimation method depends only on the first stage of the sample design and ignores other stages, then overestimation of the variance results; this should be relatively small if the first-stage sampling fraction is small. There are survey data sets, particularly government public release data, which include replication weights. These replication weights account for the design information implicitly, and SAS macros can be written to compute estimates and their standard errors fairly easily.

OBJECTIVE AND AIMS:

Our objective is to show that SAS can be used for complex survey data sets that have replication weights. In particular we aim to show that:
Aim 1: SAS can be used to calculate the mean and median and their standard errors using replicate weights.
Aim 2: Estimates will vary depending on the method of estimation used. In particular, we will compare the naïve unweighted, naïve weighted, TL and the JK2 methods.
Aim 3: Our JK2 estimates compare favorably to those produced by WesVar and SUDAAN.

BACKGROUND

DATA SOURCE

The Continuing Survey of Food Intakes by Individuals for 1994-1996, 1998 (CSFII) is a nationally representative sample of non-institutionalized persons. The CSFII was conducted by the Food Surveys Research Group, Agricultural Research Service, and United States Department of Agriculture (USDA). The CSFII is a complex multistage stratified cluster sample. Full sample weights as well as strata and primary sampling unit (PSU or Cluster) designations are included in order to do Taylor Linearization variance estimation. In addition, sets of JK2 replicate weights are available. Thus, this is an ideal publicly available data set for comparing variance estimates.

We shall present results on calcium (calcium), carbohydrates (Carbo) and total food energy (energy). For this paper, daily intakes, based on the first day's consumption (i.e. day 2 intakes excluded), will be the response variable. Respondents aged 3 and up are included in the analysis yielding 17,920 records.

According to instructions included with the CSFII public access files, the key parameters are the specification of a with-replacement (WR) design and the identification of VARSTRAT and VARUNIT as the stratum and primary sampling unit designations. Daily nutrient intakes such as what are used in this analysis are referred to as record type 40. The "read40.sas" sample program (included on the public distribution CD) with modifications was used to build the intake data set. The "jk4yrcs.sas" sample program was used to create a SAS data set containing the JK2 weights. The weight and intake data sets were merged by household ID and sample person id to form the "work.all" data set.

REPLICATION METHODS

The basic idea of replication is easy to understand. To estimate sampling errors, one repeatedly selects sub samples from the realized full sample. The desired statistics are computed from each sub sample, and the variability among these replicate estimates is used to compute the standard error of the full-sample estimate. Common replicate methods include Balanced Repeated Replication (BRR) and a variation known as the Fay method (Fay 1989), and three jackknife methods (JK1, JK2, and JKn).

Many national samples are multi-stage samples in which the first-stage units are highly stratified and two PSUs have been sampled per stratum. Thus, these designs are well suited for BRR, Fay, or JK2 with little or no modification. Simple random samples and unequal probability systematic samples are frequently easy to handle with JK1, even when the units selected are clusters rather than individual units. Random digit dialing telephone surveys using a list-assisted method (Brick et al. 1995) fit into this method. The JKn method may be appropriate for clinical trials and for establishment surveys where establishments are stratified and a different number are sampled from each stratum (Brick, Morganstein, Valliant).

Replication methods have a sound theoretical basis that allows the application of a common procedure for computation purposes, and makes feasible the calculation of standard errors for many estimators including the mean, quantiles, log odds ratios, and regression coefficients. The encoding of the design information in replicate weights rather than providing strata and PSU information may allow for improved confidentiality of the

data. Domain estimation on subsets may be problematic with TL, but can be accomplished with replication methods.
We will assume that one of the above sets of replicate weights exists. The estimated variance, $V(\hat{\theta})$, of an estimate, $\hat{\theta}$, based on the replicate estimates is

$$V(\hat{\theta}) = c\sum_{g=1}^{G} h_g(\hat{\theta}_g - \hat{\theta})^2,$$

where

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

$\hat{\theta}_g$ is the g^{th} estimate of θ based on the observations included in the g^{th} replicate,

G is the number of replicates,

c is a constant that depends on the replication method, and

h_g is a factor equal to unity for all methods except JKn, where h_g is the ratio of the number of PSUs in the stratum minus 1 to the number of PSUs. The values of *c* are given below

Method	BRR, Fay	JK1	JK2	JKn
C	1/G	|G-1| / G	1	1

SAS/MEANS
The SAS/MEANS procedure can produce unweighted and weighted means and medians . The SAS MEANS procedure will produce standard error estimates for means (weighted or unweighted) but not for medians. If the study is designed to be weighted, the unweighted 'naïve' (UWN) estimate of the mean or median will be biased. The 'naïve' weighted (WN) mean or median can be obtained using the WEIGHT statement. The standard errors for the weighted and unweighted means will be incorrect, because SAS/MEANS does not account for the stratification and clustering effects. To compute weighted quantiles, the QMETHOD=OS option is required. You can use the following code to calculate correct weighted mean and median estimates.

```
Proc means data=all QMETHOD=OS mean median print;
      var calcium carbo energy;
      weight wt4_day1;
      run;
```

SAS/SURVEYMEANS
The survey procedures were written to estimate appropriate SEs for weighted data. The SAS/SURVEYMEANS procedure does not compute weighted medians.

Weighted mean (TL)
Using the SAS SURVEYMEANS procedure, you can calculate the mean, SE, and number of observations. The variance estimation method is Taylor Linearization assuming "With Replacement". Our full sample weight is "wt4_day1" for day 1 intakes. The "strata" statement identifies the stratum indicator "varstrat" and the "cluster" statement is used to identify the PSU (or cluster) indicator "varunit." The CLASS and DOMAIN statement would be used to indicate that summaries are requested on some domain such as race as illustrated in the following code.

```
PROC SURVEYMEANS data =all;
  CLASS race;
  Domain race;
  CLUSTER varunit;
  STRATA varstrat;
   VAR calcium carbo energy;
   WEIGHT wt4_day1;
run;
```

SAS PROGRAM (JK2)
The program for the calculation of the mean and standard errors using JK2 is given below.

```
/*****************************************
  input parameters:
    indata - input data set.
    Outdata -  output data (mean and std err.)
var - analysis variable on input data set.
wt  -  variable name on input data for full sample weights.
Repwt -  prefix  used in replicate weight names. Replicate weight
names must be in the form: repwt01, repwt02, ... repwt45
n -  number of replicate weights on input data set.

Program creates an output dataset containing the method, the
number of replicates, the mean, and the standard error.
*****************************************************/

%macro jack2_mean(indata,outdata,var,wt,repwt,n);
  data rep_mean_junk2; set _null_; run; *create dummy data set*;
  proc means data=&indata noprint;
    var &var;
    weight &wt;
    output out=meanhat mean=meanhat;
  run;
  %do i=1 %to &n; ** cycle through the replicate weights **;
    %put calculating results for replicate weight &i;
     proc means data=&indata noprint;
       var &var;
       %if &i<=9 %then %do; weight &repwt.0&i; %end;
       %if &i>9 %then %do;  weight &repwt.&i; %end;
       output out=rep_mean_junk mean=mean;
     run;
    data rep_mean_junk2; set rep_mean_junk2
rep_mean_junk(in=j);
              if j then rep=&i;
    run;
  %end;

  data rep_mean_junk2; merge rep_mean_junk2 meanhat;
    by _type_;
    mean2=(mean-meanhat)*(mean-meanhat);
  run;
  proc means data=rep_mean_junk2 noprint;
    var mean2;
              id meanhat;
              output out=&outdata sum=sum;
  run;
  data &outdata; set &outdata(keep=sum meanhat);
    _METHOD_ = "JK2";
              _REPLICATES_=&n;
    &var._mean = meanhat;
              &var._std = sqrt(sum);;
              drop sum meanhat;
  run;
  options notes;
%mend jack2_mean;

%jack2_mean(all,calc_mean ,calcium,wt4_day1,r4_d1_,43);
```

We use the Woodruff method (Särndal, Swensson, and Wretman 1992) for calculating the median. It indirectly estimates the standard error of a quantile by first calculating a confidence interval on the quantile from a cumulative density function (CDF) and then using the width of the confidence interval to derive a standard error estimate. Thus, the standard error is not estimated directly from the variation of replicate quantile estimates. This is the same as the no-group method implemented by WesVar. Sudaan uses a slightly different method based on histograms based on the CDF and truncates outliers. The program for the calculation of the median and standard errors using JK2 is given below.

```
/*********************************************************
p - an integer indicating the requested quantile.  50 corresponds
to the median.
```

```
This program creates a dataset containing the median, standard
error of the median, the method used (JK2), and the number of
replicates.
*********************************************************/
%macro jack2_p(indata,outdata,var,wt,repwt,n,p);
  %let cip=.975; *For 95% CI, upper limit corresponds to 97.5% *;
  proc means data=&indata sum noprint;
    var &wt;
    output out=sumwts sum=sumwts;
  run;
  proc sort data=&indata out=rep_p_junk;
    by &var;
  run;
  data rep_p_junk; set rep_p_junk;
    _TYPE_ = 0;
  run;
  data rep_p_junk; merge rep_p_junk sumwts(drop=_FREQ_);
    retain F 0;
    by _TYPE_;
    F=F+&wt*100/sumwts;
    indic = (F<=&p);
              drop sumwts;
  run;

%jack2_mean(rep_p_junk,rep_p_junk2,indic,&wt,&repwt,&n);
  data rep_p_junk2; set rep_p_junk2;
    plo=&p-tinv(&cip,&n)*indic_std*100;
              phi=&p+tinv(&cip,&n)*indic_std*100;
    drop indic_mean indic_std;
              _TYPE_ = 0;
  run;
  data rep_p_junk3; merge rep_p_junk rep_p_junk2;
              by _TYPE_;
    retain lastF lastX xlo xp xhi;
    if _n_ > 1 then do;
      if lastF <= plo < F then Xlo=(lastX+(plo-lastF)*(&var-lastX)/(F-
lastF));
      if lastF <= &p  < F then Xp=(lastX+(&p  -lastF)*(&var-
lastX)/(F-lastF));
      if lastF <= phi < F then Xhi=(lastX+(phi-lastF)*(&var-lastX)/(F-
lastF));
              end;
              lastF = F;
              lastX = &var;
    keep xhi xlo xp;
    if last._type_ then output;
  run;
  data &outdata; set rep_p_junk3;
    _METHOD_= "JK2";
              _REPLICATES_= &n;
              &var._p&p= Xp;
              &var._p&p._std= (Xhi-Xlo)/(2*tinv(&cip,&n));
              drop xhi xlo xp;
  run;
%mend jack2_p;
%jack2_p(all,calc_med ,calcium,wt4_day1,r4_d1_,43,50);
```

SUDAAN PROGRAM

Taylor Linearization (TL)

Below is the program to calculate mean, standard error of mean, median, standard error of median, and number of observations using the SAS callable SUDAAN 8.1. The "design=WR" statement indicates that the variance estimation method is TL assuming "With Replacement". The full sample weight is "wt4_day1" for day 1 intakes. The nest statement contains the stratum indicator "varstrat" and the PSU indicator "varunit." Two calls to SUDAAN are required to get statistics on the MEAN and MEDIAN. The first proc descript is used to calculate the mean. The second is used to calculate the median. SUDAAN requires that all variables referenced be numeric. SUDAAN requires a two-part SAS dataset name "work.all" rather than just "all". SUDAAN expects the data set to be sorted by the variables on the nest statement. The following code will give estimates by race as well as overall.

```
Proc descript data =work.all design = wr;
    nest varstrat varunit;
      Weight wt4_day1;
      Var calcium carbo energy;
    subgroup r_race _one_;
    levels 3 1;
    Tables race _one_;
    Print nsum=n mean semean   /style=nchs;
    Run;

/* The default method no group method was used */

    Proc descript data =work.rt40 design = wr;
      Nest varstrat varunit;
      Weight wt4_day1;
    subgroup r_race _one_;
    levels 3 1;
    Tables race _one_;
     Var calcium carbo tenergy;
      Percentile / median;
      Tables race;
    Print nsum=n qtile=median seqtile=SE_median
    /style=nchs;
    Run;
```

Jackknife 2 (JK2)

The following code was used to calculate the mean, standard error of mean, median, standard error of median, and number of observations using the SAS callable SUDAAN 8.1. The "design=jackknife" statement indicates that the variance estimation method is Jackknife. The jackknife weights are indicated on the "jackwgts" statement. The "adjjack=1" option is used for jackknife type II. The full sample weight is "wt4_day1" for day 1 intakes. Two calls to SUDAAN are required to get statistics on the MEAN and MEDIAN. The "_one_" variable is an automatic variable that can be used for a full sample estimate.

```
    Proc descript data =work.all  design=jackknife;
      Weight wt4_day1;
    subgroup _one_;
    levels  1;
    Tables  _one_;
      Jackwgts r4_d1_01-r4_d1_43 /adjjack=1;
      Var calcium carbo energy;
    Print nsum=n mean semean   /style=nchs;
    Run;

    proc descript data =work.all design=jackknife;
      weight wt4_day1;
      jackwgts r4_d1_01-r4_d1_43 /adjjack=1;
    subgroup _one_;
    levels 1;
    Tables  _one_;
      Var calcium carbo energy;
     Percentile / median;
    Print nsum=n qtile=median seqtile=SE_median
    /style=nchs;
    Run;
```

WESVAR (JK2)

WesVar has a point and click interface. Therefore we do not give the program but only report the results for comparative purposes. The "No group" method was requested for calculating medians.

RESULTS

The estimated values of the mean and standard error (SE) and the median and SE are presented in Tables 1 and 2 respectively for the unweighted naïve (UWN), weighted naïve (WN), Taylor linearization (TL) and Jackknife 2 (JK2).

- All but the unweighted mean are the same. You can use SAS/MEANS to properly compute mean and median estimates.
- The SAS/Macro code and SUDAAN calculations are identical for the SEs for the means.
- The Jackknife estimates for standard errors are smaller than the Taylor Linearization estimates. This may not always be true.
- The Jackknife estimates from WesVar and the SAS/Macro are very close and reasonably close to those for SUDAAN.
- The SUDAAN estimates of the SE for medians are lower possibly because of the elimination of outliers.

With a similar algorithm replication estimates could be calculated to log odds ratios and linear estimates of parameters and their SEs.

CONCLUSION

It can be seen that this approach allows a wider use of SAS in complex surveys.

TABLE 1: Estimates of Day 1 MEAN Intakes

Nutrient	Software	Method	Mean	S.E.
Calcium	SAS/MEANS	UWN	810.85	3.82
(mg)	SAS/MEANS	WN	817.22	3.86
	SAS/SURVEYMEANS	TL	817.22	11.03
	SUDAAN /Descript	TL	817.22	11.03
	SUDAAN /Descript	JK2	817.22	7.39
	WesVar	JK2	817.22	7.39
	SAS / Macro	JK2	817.22	7.39
Carbo-	SAS/MEANS	UWN	247.90	0.90
hydrate	SAS/MEANS	WN	249.59	0.92
(g)	SAS/SURVEYMEANS	TI	249.59	3.07
	SUDAAN /Descript	TL	249.59	3.07
	SUDAAN /Descript	JK2	249.59	1.87
	WesVar	JK2	249.59	1.87
	SAS / Macro	JK2	249.59	1.87
Food	MEANS	UWN	1918.79	6.78
Energy	MEANS	WN	1930.26	6.88
(kcal)	SURVEYMEANS	TL	1930.26	22.85
	SUDAAN	TL	1930.26	22.85
	SUDAAN	JK2	1930.26	14.05
	WesVar	JK2	1930.26	14.05
	SAS/Macro	JK2	1930.26	14.05

Table 2: Median/SE estimates (Ungrouped Option in WesVar & SUDAAN)

Nutrient	Software	Method	Median	S.E.
Calcium	SAS/MEANS	UWN	714.86	N/A
(mg)	SAS/MEANS	WN	718.94	N/A
	SAS/SURVEYMEANS	TL	N/A	N/A
	SUDAAN /Descript	TL	718.92	11.49
	SUDAAN /Descript	JK2	718.92	7.99
	WesVar	JK2	718.92	7.80
	SAS / Macro	JK2	718.92	7.86
Carbo-	SAS/MEANS	UWN	228.03	N/A
hydrate	SAS/MEANS	WN	228.62	N/A
(g)	SAS/SURVEYMEANS	TI	N/A	N/A
	SUDAAN /Descript	TL	228.61	2.17
	SUDAAN /Descript	JK2	228.61	1.65
	WesVar	JK2	228.61	1.49
	SAS / Macro	JK2	228.612	1.49
Food	MEANS	UWN	1746.75	N/A
Energy	MEANS	WN	1754.63	N/A
(kcal)	SURVEYMEANS	TI	N/A	N/A
	SUDAAN	TL	1754.62	14.79
	SUDAAN	JK2	1754.62	9.31
	WesVar	JK2	1754.62	11.26
	SAS/Macro	JK2	1754.62	11.26

REFERENCES

Brick, J.M., and Morganstein, D. (1996). WesVarPC: Software for Computing VarianceEstimates from Complex Designs. *Proceedings of the 1996 Annual Research Conferenc*e, pp. 861-866. Washington, DC: U.S. Bureau of the Census.

"Analysis of Complex Samples Using Replication" by J. Michael Brick, David Morganstein, and Richard Valliant http://www.westat.com/wesvar/techpapers/index.html

Brick, J.M., and Kalton, G. (1996). Handling Missing Data in Survey Research. *Statistical Methods in Medical Research,* 5, 215-238.

Brick, J.M., and Morganstein, D. (1997). Computing Sampling Errors from Clustered Unequally Weighted Data Using Replication: WesVarPC. *Bulletin of the International Statistical Institut*e, *Proceeding*s, Book 1, pp. 479-482.

Brick, J.M., Waksberg, J., Kulp, D., and Starer, A. (1995). Bias in List-Assisted Telephone Surveys. *Public Opinion Quarterl*y, **5**9(2), 218-235.

Brogan, D.J. (1998). Pitfalls of Using Standard Statistical Software Packages for Sample Survey Data. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics.* New York: John Wiley and Sons.

Cochran, W.G. (1977). *Sampling Techniques.* New York: John Wiley and Sons.

DiGaetano, R., Brick, J.M., and Flores-Cervantes, I. (1998). Preserving Degrees of Freedom in a Multi-mode, Multi-site Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Associatio*n, pp. 475-480.
Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plan*s. Philadelphia:Society for Industrial and Applied Mathematics.

Fay, R.E. (1989). Theoretical Application of Weighting for Variance Calculation. *Proceedings of the Section on Survey Research Methods of the American Statistical Associatio*n, pp. 212-217.

Graubard, B.I., and Korn, E.L. (1996). Survey Inference for Subpopulations. *American Journal of Epidemiolog*y, **14**4, 102-106.

Hinkins, S., Moriarity, C., and Scheuren, F. (1996). Replicate Variance Estimation in Stratified Sampling with Permanent Random Numbers. *Proceedings of the Section on SurveyResearch Methods of the American Statistical Associatio*n, pp. 824-829.

Kish, L. (1992). Weighting for Unequal Pi. *Journal of Official Statistic*s, 8, 183-200.

Kish, L., and Frankel, M. (1974). Inference from Complex Samples. *Journal of the Royal Statistical Society* B, **3**6, 1-22.

Korn, E.L., and Graubard, B.I. (1995). Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. *The American Statistician*, **4**9, 291-295.

Kovar, J.G., Rao, J.N.K., and Wu, C.F.J. (1988). Bootstrap and Other Methods to Measure Errors in Survey Estimates. *Canadian Journal of Statistic*s, **1**6, 25-46.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of Linearization, Jackknife and Balanced Repeated Replication Methods. *Annals of Statistic*s, 9, 1010-1019.

McCarthy, P.J. (1966). Replication: An Approach to the Analysis of Data from Complex Surveys. *Vital and Health Statistic*s, Series 2, No. 14. Hyattsville, MD: National Center for Health Statistics.

SUDAAN® (2001). SUDAAN VERSION 8.1:Software for the Statistical Analysis of Correlated Data. Research Triangle Institute. NC.

U.S. Department of Agriculture, Agricultural Research Service. 2000. Continuing Survey of Food Intakes by Individuals 1994-96, 1998. CD-ROM.

WesVar®(2001). WesVar is a trademark of Westat for its proprietary computer software. 1650 Research Boulevard, Rockville, MD 20850, USA. Telephone 1-800-WESTAT1.

ACKNOWLEDGMENTS

This study was partially funded by Agricultural Research Service, United States Department of Agriculture, Project No. 6251-53000-002-00D, the Delta NIRI project.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. The authors also have versions of the macros for doing domain analysis. Contact the author at:

Jeff Gossett
800 Marshall St Slot 512-26
Little Rock, AR 72202
Work Phone: (501) 320-6631
Fax: (501) 320-1552
Email: GossettJeffreyM@uams.edu