

Paper 367-2008

## Try, Try Again: Replication-Based Variance Estimation Methods for Survey Data Analysis in SAS® 9.2

Pushpal K Mukhopadhyay, Anthony B. An, Randall D. Tobias, and Donna L. Watts  
SAS Institute Inc., Cary, NC

### ABSTRACT

Complex survey samples are constructed with selection schemes that affect the usual random assumptions, so SAS/STAT® software provides specialized procedures to analyze them: SURVEYMEANS, SURVEYFREQ, SURVEYREG, and SURVEYLOGISTIC for means, frequencies, regression, and logistic analysis, respectively. These procedures all use the Taylor series expansion method for variance estimation, which is usually considered to be the "gold standard" when it is practical to compute. However, replication methods are also widely used in practice for variance estimation. Replication methods, such as the jackknife and balanced repeated replication (BRR), replace complex algebra with simple repeated analysis. They enable you to analyze the data without the original sample design, protecting survey security, and they ease the task of estimating variances for nonlinear quantities.

With the release of SAS 9.2, the SAS/STAT survey analysis procedures now also implement replication methods. These include standard approaches such as jackknife and BRR as well as customized replication methods that employ user-supplied replicate weights. This paper discusses replication methods, comparing them to the Taylor series expansion method with respect to both technical characteristics and practical utility. This paper also discusses other significant enhancements to the survey design and analysis procedures in SAS 9.2.

### INTRODUCTION

Sample surveys provide information about a finite population by observing only a fraction of the population. To provide statistically valid inference, samples are typically selected through randomization techniques. Statistical agencies such as the U.S. Census Bureau, Bureau of Labor Statistics, Statistics Canada, and Statistics Sweden conduct surveys to collect information about social and economic conditions of people, households, businesses, and industries. Surveys of natural resources and opinion polls are also common. Most of these surveys collect data through complex designs that include stratification and clustering, requiring special techniques for analysis; see Lohr (1999), Särndal, Swenson, and Wretman (1992), and Chamber and Skinner (2003). SAS/STAT software provides specialized procedures to analyze survey data: SURVEYMEANS, SURVEYFREQ, SURVEYREG, and SURVEYLOGISTIC for means, frequencies, regression, and logistic analysis, respectively.

The most commonly used variance estimation method for survey data is the Taylor series expansion method. This method obtains a linear approximation of an estimator by using a Taylor series expansion. The precision of the linearized statistic is then estimated by using standard survey variance estimation methods. Taylor series expansion is often considered to be the "gold standard" for survey variance estimation, but it can be complicated to derive for some estimators that are nonlinear functions of means. It also requires you to specify variables that contain strata and cluster information (for a stratified multistage design). Strata or cluster information might not be available due to data confidentiality. This identification requirement can be a major limitation for the Taylor series expansion method.

An alternative to the Taylor series expansion method that addresses both the complexity issue and the identification requirement is replication-based variance estimation. A replicate sample is a subsample of the original sample. An estimate of a quantity of interest is obtained for each replicate sample. The variability of the estimated quantity among the replicate samples is then used as a replication-based estimator of variance. In order to obtain a statistically justified estimator of variance, each replicate sample should be drawn by following some specific resampling scheme. Currently SAS/STAT survey analysis procedures support the two most widely used replication variance estimation methods, the jackknife and balanced repeated replication (BRR).

The SURVEYMEANS, SURVEYFREQ, SURVEYREG, and SURVEYLOGISTIC procedures now support the Taylor series expansion, jackknife, and BRR variance estimation, by using the PROC statement options `VARMETHOD = TAYLOR`, `VARMETHOD=JACKKNIFE`, and `VARMETHOD=BRR`, respectively. This paper illustrates the application of different variance estimation methods by using data collected through a complex national survey.

## THE JACKKNIFE VARIANCE ESTIMATION METHOD

The jackknife variance estimation method in SAS is available for any survey design. For simplicity, consider a clustered stratified sample design where the first stage clusters, or primary sampling units (PSUs), are selected by using a simple random sample with replacement. Assume  $n_h$  sample PSUs are selected from  $N_h$  population PSUs in stratum  $h$ , where  $h = 1, 2, \dots, H$ . Let  $\theta$  be a finite population quantity of interest and  $\hat{\theta}$  be a sample-based estimator of  $\theta$ . The *delete-1 jackknife* method deletes one PSU at a time and adjusts the full sample weight for the other clusters in that stratum, repeating the process for each stratum independently. The adjusted observation weights in each replicate sample are called replicate weights. Let  $\hat{\theta}_r$  denote the estimate of  $\theta$  obtained from the  $r$ th replicate weights. Then the jackknife variance estimator of  $\hat{\theta}$  is

$$\hat{V}(\hat{\theta}) = \sum_{r=1}^R \alpha_r (\hat{\theta}_r - \hat{\theta})^2 \quad (1)$$

where  $\alpha_r = n_h^{-1}(n_h - 1)$  and  $R$  is the total number replicates. The quantity  $\alpha_r$  is also called the jackknife coefficient. In this example, the total number of replicates is the same as the total number of clusters in the full sample. See Wolter (1985), Rust (1985), and Shao and Tu (1995) for details. Unless specified otherwise, the term 'jackknife' method denotes the delete-1 jackknife method throughout this paper.

## THE BALANCED REPEATED REPLICATION METHOD

Another replication-based variance estimation method is *balanced repeated replication* (BRR). The most common form of the BRR method is suitable for sample designs that have a large number of strata with two PSUs in each stratum, where the PSUs are selected with replacement. A replicate sample (also known as half sample) is obtained by deleting one PSU per stratum and doubling the original weight of the remaining PSU. To satisfy certain balance conditions, the PSUs are deleted according to a corresponding Hadamard matrix. See Wolter (1985) for an introduction to and construction of Hadamard matrices. The BRR variance estimator of a full sample estimator  $\hat{\theta}$  is given by

$$\hat{V}(\hat{\theta}) = R^{-1} \sum_{r=1}^R (\hat{\theta}_r - \hat{\theta})^2 \quad (2)$$

where  $\hat{\theta}_r$  is an estimator of  $\theta$  using the  $r$ th balanced half sample and  $R$  is the total number of replicates. See Wolter (1985), Rust (1985), and Shao and Tu (1995) for details.

In many situations, especially for nonlinear estimators, one or more replicate estimators  $\hat{\theta}_r$  might be undefined but the full sample estimator  $\hat{\theta}$  is defined. Fay's BRR method adjusts the original weight by a coefficient  $\epsilon$  ( $0 \leq \epsilon < 1$ ) so that the replicate estimators are defined for all replicate samples. This method is similar to the traditional BRR method, but instead of deleting one PSU per stratum, it multiplies the original weight by the coefficient  $\epsilon$ . The original weight for the remaining PSU in that stratum is multiplied by  $2 - \epsilon$ . The Fay's BRR variance estimator of  $\hat{\theta}$  is computed as

$$\hat{V}(\hat{\theta}) = \left\{ R(1 - \epsilon)^2 \right\}^{-1} \sum_{r=1}^R (\hat{\theta}_r - \hat{\theta})^2 \quad (3)$$

where  $0 \leq \epsilon < 1$ . See Dippo, Fay, and Morganstein (1984), Fay (1989), Judkins (1990), and Rao and Shao (1999) for more information. Note that when  $\epsilon = 0$ , then Fay's BRR method becomes the traditional BRR method.

## SYNTAX FOR REPLICATION METHODS

The new syntax for specifying replication-based variance estimation consists of the VARMETHOD= option in the PROC statement:

```
VARMETHOD=BRR
VARMETHOD=JACKKNIFE | JK
VARMETHOD=TAYLOR
```

and the REPWEIGHTS statement for specifying user-defined replication weights:

```
REPWEIGHTS variables;
```

These options and statement are available in all four survey analysis procedures. The VARMETHOD=BRR and VARMETHOD=JACKKNIFE options and the REPWEIGHTS statement have further sub-options, whose complete details are described in the following chapters of the *SAS/STAT User's Guide*: "The SURVEYMEANS Procedure," "The SURVEYFREQ Procedure," "The SURVEYREG Procedure," and "The SURVEYLOGISTIC Procedure."

## THE MEPS SURVEY

The Medical Expenditure Panel Survey (MEPS) is a nationally representative survey of U.S. civilian noninstitutionalized population, conducted annually by the U.S. Department of Health and Human Resources. The main objectives are to determine the cost of specific health services and the quality of health insurance available to U.S. workers. The household component (HC) of the MEPS collects demographic characteristics, health conditions, health insurance coverage, and health care expenditure information through household interviews. Data are collected by using a sample of families and individuals through a multistage overlapping panel design. The primary sampling units are defined by geographic locations such as counties, small groups of counties, or metropolitan statistical areas. Within a PSU, area segments and permit area segments are used as second stage units. See the Web site at [http://www.meps.ahrq.gov/mepsweb/about\\_meps/survey\\_back.jsp](http://www.meps.ahrq.gov/mepsweb/about_meps/survey_back.jsp) for a detailed description of the survey design. The 1999 full-year consolidated data file contains 24,618 individuals who are divided into 143 strata and 460 PSUs. The strata and PSU information are useful for variance estimation purposes. The 1999 full-year consolidated data file HC-038 (MEPS HC-038, 2002) from the MEPS is used to illustrate different variance estimation methods that are available in PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, and PROC SURVEYLOGISTIC procedures. The data can be downloaded directly from the Agency for Healthcare Research and Quality (AHRQ) Web site at [http://www.meps.ahrq.gov/mepsweb/data\\_stats/download\\_data\\_files\\_detail.jsp?choPufNumber=HC-038](http://www.meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?choPufNumber=HC-038) in either ASCII format or SAS transport format.

For the examples used in this paper, the analysis variables are the following individual level items for 1999:

- **expenditure**—total health care expenditure
- **insuranceType**—type of insurance coverage
- **poverty**—poverty category
- **totalIncome**—total income

In addition, the following demographic variables are used as covariates and for classifications:

- **sex**—gender
- **age99x**—age
- **region99**—census region of residence

Finally, the following variables are used for design specifications:

- **varianceStrata**—strata identification
- **variancePSU**—PSU identification
- **perwt99**—person level weights

The following SAS statements create a data set for illustration. The input data set `meps.h38` has been downloaded from the MEPS website. Note that this data set is used only to demonstrate different variance estimation methods available in SAS/STAT 9.2. Neither the data set nor the SAS statements should be used for inferential purposes.

```

libname meps '';
data exempledata;
  set meps.h38;
  age = age99x;
  if age99x = -1 then do;
    if age42x = -1 then age = age31x;
    else age = age42x;
  end;
  region = region99;
  if region99 = -1 then do;
    if region42 = -1 then region = region31;
    else region = region42;
  end;
  totexpnnonzero = expenditure + 1;
  logExpenditure = log(totexpnnonzero);
  if inscov99 = 3 then insured = 'NO';
  else insured = 'YES';
  rename totexp99 = expenditure      inscov99 = insuranceType
         povcat99 = poverty          ttlp99x  = totalIncome
         varstr99 = varianceStrata  varpsu99 = variancePSU
         perwt99f = personWeight;
run;

```

## THE JACKKNIFE VARIANCE ESTIMATION METHOD

### SYNTAX

Use the `VARMETHOD = JACKKNIFE | JK < method-options >` option in the `PROC` statement to request the jackknife variance estimation method.

You can specify the following *method-options* in parentheses after the `VARMETHOD=JACKKNIFE` option:

#### **OUTJKCOEFS**=SAS-data-set

names a SAS data set to store the jackknife coefficients.

#### **OUTWEIGHTS**=SAS-data-set

names a SAS data set to store the replicate weights that the procedure creates for jackknife variance estimation.

The `OUTWEIGHTS=` method-option is not available when you provide replicate weights with a `REPWEIGHTS` statement.

You can specify the optional `STRATA` or `CLUSTER` statement with the `VARMETHOD = JACKKNIFE` option. The only requirement is at least two PSUs/observations per stratum for a stratified design or at least two PSUs/observations in the data set. The `JACKKNIFE` syntax is identical for all procedures. For details, see the following chapters of the *SAS/STAT User's Guide*: "The SURVEYMEANS Procedure," "The SURVEYFREQ Procedure," "The SURVEYREG Procedure," and "The SURVEYLOGISTIC Procedure."

## ESTIMATES FOR POPULATION MEANS

Suppose you want to estimate the mean of total health care expenditure of a person for the 1999 population. You can use the SURVEYMEANS procedure with `expenditure` as the analysis variable. The following SAS statements use the Taylor series expansion method to estimate the variance of the estimated mean. The `STRATA` statement specifies the stratification variable, the `CLUSTER` statement specifies the PSU identification, and the `WEIGHT` statement specifies the individual level survey weights. The `VARMETHOD = TAYLOR` option in the `PROC` statement requests the Taylor series expansion method.

```

proc surveymeans data = exempledata varmethod = taylor;
  strata varianceStrata;
  cluster variancePSU;
  weight personWeight;
  var expenditure;
run;

```

Figure 1 displays the data summary and estimated values produced by PROC SURVEYMEANS. 1053 observations with nonpositive weights are not used for the subsequent analysis. The estimated mean health care expenditure using 23565 individuals is 2156.47 with a standard error of 62.72. The 95% confidence interval for the mean expenditure is (2033.06, 2279.88).

Figure 1 The Taylor Series Expansion Method for the SURVEYMEANS Procedure

The SURVEYMEANS Procedure					
Data Summary					
	Number of Strata				143
	Number of Clusters				460
	Number of Observations				24618
	Number of Observations Used				23565
	Number of Obs with Nonpositive Weights				1053
	Sum of Weights				276410767
Statistics					
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean
expenditure	TOTAL HEALTH CARE EXP 99	23565	2156.468447	62.723462	2033.06156 2279.87533

To estimate the variance of the estimated mean by using the jackknife variance estimation method, you simply need to specify `VARMETHOD = JACKKNIFE` as a PROC statement option. The following SAS statements obtain an estimate of the variance by using the jackknife variance estimation method.

```
proc surveymeans data = exampledata varmethod = jackknife;
  strata varianceStrata;
  cluster variancePSU;
  weight personWeight;
  var expenditure;
run;
```

There are a total of 460 PSUs. The jackknife variance estimation method creates 460 replicate samples by deleting one PSU at a time. Since most surveys contain a large number of PSUs, it is computationally efficient to save the replicate weights and the jackknife coefficients in SAS data sets and use the saved values for subsequent analyses. The replicate weights and the jackknife coefficients can be saved in SAS data sets by using the `OUTWEIGHTS =` and the `OUTJKCOEFS =` method-options for `VARMETHOD = JACKKNIFE`. The following SAS statements use the jackknife variance estimation method and save replicate weights and jackknife coefficients in SAS data sets `jkrepweights` and `jkcoefficients`, respectively. The data set `jkrepweights` contains all the variables in the data set `exampledata`, in addition to the replicate weight variables named `RepWt_1` to `RepWt_460`.

```
proc surveymeans data = exampledata
  varmethod = jackknife (outweights = jkrepweights
    outjkcoefs = jkcoefficients);
  strata varianceStrata;
  cluster variancePSU;
  weight personWeight;
  var expenditure;
run;
```

Figure 2 displays the variance estimation method and Figure 3 displays the estimated values produced by PROC SURVEYMEANS. There are a total of 460 replicates generated by the procedure. The estimated mean health care expenditure is 2156.47 with a standard error of 62.74. The 95% confidence interval for the mean expenditure is (2033.03, 2279.90).

**Figure 2** The Jackknife Variance Estimation Method for the SURVEYMEANS Procedure

The SURVEYMEANS Procedure	
Variance Estimation	
Method	Jackknife
Number of Replicates	460

**Figure 3** The Jackknife Variance Estimation Method for the SURVEYMEANS Procedure, Estimated Values

Statistics					
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean
expenditure	TOTAL HEALTH CARE EXP 99	23565	2156.468447	62.737868	2033.03322 2279.90367

### ESTIMATES FOR A CONTINGENCY TABLE

Suppose you want to estimate the percentage of individuals within each category of health insurance coverage and the percentage of individuals within each cross-classification of poverty categories and insurance categories for the 1999 population. You can use the SURVEYFREQ procedure with insuranceType and povact99 as analysis variables. The following SAS statements request a two-way table for insurance coverage by poverty categories and use the jackknife variance estimation method to calculate standard errors.

```
proc surveyfreq data = jkrepweights varmethod = jackknife;
  weight personWeight;
  repweights RepWt_1-RepWt_460 / jkcoefs = jkcoefficients;
  tables insuranceType*povact99;
run;
```

The data set jkrepweights obtained from the previous PROC SURVEYMEANS statements contains all the variables in the data set exampladata, in addition to the replicate weight variables RepWt\_1 to RepWt\_460. The REPWEIGHTS statement specifies the names for the replicate weight variables. The REPWEIGHTS statement option JKCOEFS specifies the SAS data set that contains the jackknife coefficient for each observation. In this particular example, the SURVEYFREQ procedure will not generate the replicate weights and the jackknife coefficients. The input replicate weights and jackknife coefficients contain all the necessary information for variance estimation, so that if the replicate weights and the jackknife coefficients are specified, then the strata and the cluster identifications are not required. The VARMETHOD = JACKKNIFE option in the PROC statement requests the jackknife variance estimation method. Figure 4 displays the two-way table of insuranceType and povact99. Standard errors of estimated quantities are calculated by using the jackknife method.

Figure 4 The Jackknife Variance Estimation Method for the SURVEYFREQ Procedure

The SURVEYFREQ Procedure						
Table of insuranceType by poverty						
insuranceType	poverty	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
1	1	936	9689015	624759	3.5053	0.2066
	2	498	5032074	464129	1.8205	0.1649
	3	2040	22442179	1129017	8.1191	0.3302
	4	5857	69835693	2816150	25.2652	0.6174
	5	6799	97405036	3721701	35.2392	0.7943
	Total	16130	204403997	6554617	73.9494	0.6889
2	1	1920	16708348	1034325	6.0448	0.3676
	2	487	4581804	420803	1.6576	0.1487
	3	844	9183636	574957	3.3225	0.2125
	4	672	7758261	486381	2.8068	0.1815
	5	318	3577523	284380	1.2943	0.0994
	Total	4241	41809572	1559405	15.1259	0.5543
3	1	806	6397775	500629	2.3146	0.1714
	2	338	2708118	270952	0.9797	0.0950
	3	773	7302179	503540	2.6418	0.1716
	4	877	8587833	575319	3.1069	0.1734
	5	400	5201293	380308	1.8817	0.1284
	Total	3194	30197198	1319234	10.9248	0.3587
Total	1	3662	32795137	1527870	11.8646	0.5000
	2	1323	12321996	774817	4.4579	0.2656
	3	3657	38927994	1536394	14.0834	0.4435
	4	7406	86181788	3269175	31.1789	0.6907
	5	7517	106183852	3946498	38.4152	0.8241
	Total	23565	276410767	7790639	100.000	

## ESTIMATES FOR REGRESSION COEFFICIENTS

Suppose you want to estimate the regression coefficients for the 1999 population when the natural log of total health care expenditure (`logExpenditure`) is regressed on age (`age`) and gender (`sex`). You can use the `SURVEYREG` procedure with `logExpenditure` as the dependent variable, as shown in the following statements.

```
proc surveyreg data = exampladata varmethod = taylor;
  strata varianceStrata;
  cluster variancePSU;
  weight personWeight;
  class sex;
  model logExpenditure = age sex / solution;
run;
```

The `STRATA` statement specifies the stratum identification, the `CLUSTER` statement specifies the PSU identification, the `WEIGHT` statement specifies individual level survey weights, and the `CLASS` statement specifies the classification variable `sex`. The dependent variable and the independent variables are specified in the `MODEL` statement. The `SOLUTION` option in the `MODEL` statement displays the parameter estimates. The standard errors of the estimated regression coefficients are calculated by using the Taylor series expansion method. The `VARMETHOD = TAYLOR` option in the `PROC` statement requests the Taylor series expansion method. Figure 5 displays the parameter estimates along with their standard errors produced by PROC SURVEYREG.

**Figure 5** The Taylor Series Expansion Method for the SURVEYREG Procedure

The SURVEYREG Procedure				
Regression Analysis for Dependent Variable logExpenditure				
Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	4.5501064	0.05728676	79.43	<.0001
age	0.0376353	0.00096827	38.87	<.0001
SEX 1	-0.7890617	0.03962576	-19.91	<.0001
SEX 2	0.0000000	0.00000000	.	.

NOTE: The denominator degrees of freedom for the t tests is 317.  
Matrix X'WX is singular and a generalized inverse was used to solve the normal equations. Estimates are not unique.

To estimate the variance of estimated regression coefficients by using the jackknife method, you can again use the replicate weights and the jackknife coefficients obtained from the previous SURVEYMEANS procedure. The following SAS statements invoke the jackknife variance estimation method.

```
proc surveyreg data = jkrepweights varmethod = jackknife;
  weight personWeight;
  class sex;
  repweights RepWt_1-RepWt_460 / jkcoefs = jkcoefficients;
  model logExpenditure = age sex / solution;
run;
```

The data set jkrepweights obtained from the previous PROC SURVEYMEANS statements contains all the variables in the data set exampladata, in addition to the replicate weight variables RepWt\_1 to RepWt\_460. The **REPWEIGHTS** statement specifies the names for the replicate weight variables. The **REPWEIGHTS** statement option **JKCOEFS** specifies the SAS data set that contains the jackknife coefficients for each observation. The **VARMETHOD = JACKKNIFE** option in the **PROC** statement requests the jackknife variance estimation method. [Figure 6](#) displays the parameter estimates along with their standard errors produced by PROC SURVEYREG. The parameter estimates from [Figure 6](#) match the parameter estimates from [Figure 5](#). The standard errors from [Figure 6](#) are little higher than the standard errors from [Figure 5](#).

**Figure 6** The Jackknife Variance Estimation Method for the SURVEYREG Procedure

The SURVEYREG Procedure				
Regression Analysis for Dependent Variable logExpenditure				
Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	4.5501064	0.05730160	79.41	<.0001
age	0.0376353	0.00096851	38.86	<.0001
SEX 1	-0.7890617	0.03963330	-19.91	<.0001
SEX 2	0.0000000	0.00000000	.	.

NOTE: The denominator degrees of freedom for the t tests is 460.  
Matrix X'WX is singular and a generalized inverse was used to solve the normal equations. Estimates are not unique.

## ESTIMATES FOR LOGISTIC REGRESSION COEFFICIENTS

To estimate the logistic regression coefficients for the 1999 population when the categorical variable type of insurance coverage (insuranceType) is regressed on total individual income (totalIncome) and census region (region), you can use the SURVEYLOGISTIC procedure. The following SAS statements invoke the jackknife variance estimation method in PROC SURVEYLOGISTIC.



```

proc surveylogistic data = jkrepweights varmethod = jackknife;
  weight personWeight;
  class region;
  model insuranceType = totalIncome region / link = glogit;
  repweights RepWt_1-RepWt_460 / jkcoefs = jkcoefficients;
run;

```

The data set `jkrepweights` obtained from the previous PROC SURVEYMEANS statements is used. The `LINK = GLOGIT` option in the model statement specifies a generalized logit link. The `REPWEIGHTS` statement specifies the names for the replicate weight variables. The `REPWEIGHTS` statement option `JKCOEFS` specifies the SAS data set that contains the jackknife coefficients for each observation. The `VARMETHOD = JACKKNIFE` option in the PROC statement requests the jackknife variance estimation method. Figure 7 displays the parameter estimates along with their standard error produced by PROC SURVEYLOGISTIC.

**Figure 7** The Jackknife Variance Estimation Method for the SURVEYLOGISTIC Procedure

The SURVEYLOGISTIC Procedure						
Analysis of Maximum Likelihood Estimates						
Parameter	insurance Type	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	1.5634	0.0500	977.4799	<.0001
Intercept	2	1	0.6707	0.0571	137.7779	<.0001
totalIncome	1	1	0.000025	1.816E-6	185.7235	<.0001
totalIncome	2	1	-0.00003	2.619E-6	108.2115	<.0001
region	1 1	1	0.2282	0.0767	8.8578	0.0029
region	1 2	1	0.2961	0.0978	9.1721	0.0025
region	2 1	1	0.3654	0.0777	22.1436	<.0001
region	2 2	1	0.1189	0.0898	1.7521	0.1856
region	3 1	1	-0.3152	0.0623	25.6126	<.0001
region	3 2	1	-0.3098	0.0715	18.7626	<.0001

## THE BRR VARIANCE ESTIMATION METHOD

### SYNTAX

Use the `VARMETHOD = BRR < method-options >` option in the PROC statement to request the BRR variance estimation procedure.

The BRR method requires a stratified sample design with two primary sampling units (PSUs) in each stratum. If you specify the `VARMETHOD=BRR` option, you must also specify a `STRATA` statement unless you provide replicate weights with a `REPWEIGHTS` statement. BRR syntax is identical to all procedures.

You can specify the following *method-options* in parentheses after the `VARMETHOD=BRR` option.

#### **FAY** <=value>

requests Fay's method, which is a modification of the BRR method. You can specify the *value* of the Fay coefficient, which must be a nonnegative number less than 1. By default, the value of the Fay coefficient equals 0.5.

#### **HADAMARD**=SAS-data-set

#### **H**=SAS-data-set

names a SAS data set that contains the Hadamard matrix for BRR replicate construction. If you do not provide a Hadamard matrix with the `HADAMARD=` method-option, the procedure generates an appropriate Hadamard matrix for replicate construction.

If you do not specify the `REPS=` method-option, then the number of replicates is taken to be the number of observations in the `HADAMARD=` input data set. If you specify the number of replicates by using the `REPS = nreps` method-option, then the first *nreps* observations in the `HADAMARD=` data set are used to construct the replicates.

**OUTWEIGHTS=SAS-data-set**

names a SAS data set to store the replicate weights created by the procedure for BRR variance estimation.

The **OUTWEIGHTS=** method-option is not available when you provide replicate weights with a **REPWEIGHTS** statement.

**PRINTH**

displays the Hadamard matrix used to construct replicates for BRR. When you provide the Hadamard matrix in the **HADAMARD=** method-option, the procedure displays only the rows and columns that are actually used to construct replicates.

The **PRINTH** method-option is not available when you provide replicate weights with a **REPWEIGHTS** statement because the procedure does not use a Hadamard matrix in this case.

**REPS=nreps**

specifies the number of replicates for BRR variance estimation. The value of *nreps* must be an integer greater than 1.

If you do not provide a Hadamard matrix with the **HADAMARD=** method-option, the number of replicates should be greater than the number of strata and should be a multiple of 4. If a Hadamard matrix cannot be constructed for the value that you specify, the value is increased until a Hadamard matrix of that dimension can be constructed. Therefore, it is possible for the actual number of replicates to be larger than the value that you specify.

If you provide a Hadamard matrix with the **HADAMARD=** method-option, the value of the **REPS=** option must not be less than the number of rows in the Hadamard matrix. If you provide a Hadamard matrix and do not specify the **REPS=** method-option, the number of replicates equals the number of rows in the Hadamard matrix.

If you provide replicate weights with a **REPWEIGHTS** statement, the procedure does not use the **REPS=** method-option.

For details, see the following chapters of the *SAS/STAT User's Guide*: "The SURVEYMEANS Procedure," "The SURVEYFREQ Procedure," "The SURVEYREG Procedure," and "The SURVEYLOGISTIC Procedure."

**ESTIMATES FOR POPULATION MEANS**

The highest level of stratification is achieved when there are only two PSUs per stratum. Many surveys are stratified up to two PSUs per stratum. The BRR variance estimation method in SAS is applicable only to surveys with two PSUs per stratum design. If your survey design does not have this design, then there are several methods in the literature to overcome this restriction; see Wolter (1985) and Shao and Tu (1995). The 1999 MEPS survey has 143 strata and the observed number of PSUs per stratum ranges from 2 to 39. For simplicity, the grouped balanced half sample (GBHS) method is used to demonstrate the BRR variance estimation technique for the MEPS survey. See "[APPENDIX: BRR FOR DESIGNS WITH MORE THAN TWO PSUS PER STRATUM](#)" on page 14 for an introduction to the GBHS method. There is no theoretical reason to choose the GBHS method for the MEPS survey, but it illustrates the syntax for the BRR. The procedure is computationally convenient, but some information is lost.

The following SAS statements calculate the standard error of the estimated mean total health care expenditure of a person for the 1999 population by using the GBHS method. The SAS statements to create the dataset *brrexample* are given in "[APPENDIX: BRR FOR DESIGNS WITH MORE THAN TWO PSUS PER STRATUM](#)" on page 14.

```
proc surveymeans data = brrexample varmethod = brr;
  strata varianceStrata;
  cluster brrpsu;
  weight personWeight;
  var expenditure;
run;
```

The **STRATA** statement specifies stratum identification, the **WEIGHT** statement specifies individual level survey weights, and the **VAR** statement specifies the analysis variable *expenditure*. The **VARMETHOD = BRR** option in the **PROC** statement requests the BRR variance estimation method. The **CLUSTER** statement specifies the modified PSU identification. There are exactly two *brrpsu* per stratum in the modified data set *brrexample*. You can also use the **OUTWEIGHTS=** method-option for **VARMETHOD = BRR** to save replicate weights in a SAS data set. You can specify the **REPWEIGHTS** statement to use the saved replicate weights for subsequent analyses. There are a total of 144 replicate samples generated by using the GBHS method.

Figure 8 displays the variance estimation method and Figure 9 displays the estimated values produced by the PROC SURVEYMEANS. There are a total of 144 replicate samples by using the GBHS method. The estimated mean is 2156.47 with a standard error of 74.12. The 95% confidence interval for the mean health care expenditure is (2009.95, 2302.98). Note that the number of replicates for the GBHS method is much smaller than the number of replicates for the delete-1 jackknife method.

**Figure 8** The BRR Method for the SURVEYMEANS Procedure

The SURVEYMEANS Procedure	
Variance Estimation	
Method	BRR
Number of Replicates	144

**Figure 9** The BRR Method for the SURVEYMEANS Procedure, Estimated Values

		Statistics			
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean
expenditure	TOTAL HEALTH CARE EXP 99	23565	2156.468447	74.121505	2009.95305 2302.98385

The estimated mean health care expenditure does not depend on the variance estimation method. The mean health care expenditure is estimated as 2156.47 for all three methods. The standard errors of the estimated mean total health care expenditure are not the same for the Taylor series expansion, jackknife, and GBHS methods. For this particular example, the standard error (62.738) using the jackknife method is higher than the standard error (62.714) using the Taylor series expansion method.

## COMMENTS ON DIFFERENT VARIANCE ESTIMATION METHODS

With the survey procedures in SAS/STAT 9.2, you have the flexibility to use the Taylor series expansion, the jackknife, or the BRR variance estimation method. All three methods estimate an approximation of the true variance and have nice properties for large samples (Krewski and Rao 1981). Several investigators have compared the analytical and empirical performances of the Taylor series expansion, jackknife, and BRR variance estimation methods. A good review is given in Rust (1985). Performance of these methods depend mainly on the survey design and the choice of estimator. No one method can be recommended as the 'best' across all survey designs and for all estimators. See Kish and Frankel (1968, 1970, 1974), Krewski and Rao (1981), Rao and Wu (1983), and Kovar, Rao, and Wu (1988) for comparisons among variance estimation methods for different survey designs. For reasonably large sample sizes, all three methods perform satisfactorily. The choice between these three variance estimators depends mainly on the design of the survey, computational efficiency, and confidentiality constraints. We discuss some important aspects of using these methods in SAS.

With respect to the survey design, in principle the Taylor series expansion can be applied most often and BRR least often. A well-developed theory is available for the Taylor series expansion method for all common survey designs; the jackknife method is applicable to any designs with or without stratification or clustering; and the BRR method is applicable only for designs with two PSUs per stratum. Performance of the delete-1 jackknife method might not be satisfactory for unequal probability without replacement designs (Lohr 1999; Berger and Skinner 2005). If the sample size is not large enough, the estimate of variances when using the Taylor series expansion method is often biased downward (Lohr 1999). For many nonlinear functions of the mean, the estimated variance when using the jackknife method is often larger than the estimated variance when using the Taylor series expansion method (Fuller 2006).

In the technical literature, the methods are often distinguished by the relative difficulties of implementing them. The Taylor series expansion method requires a separate variance formula for each nonlinear statistic. In contrast, the replication methods use the same variance formula for every statistic. This distinction is less of an issue in SAS. Since SAS/STAT survey analysis procedures provide Taylor series expansion methods for all well-known survey estimators, you do not need to compute any variance formulas for these estimators. The number of replicate samples for the BRR method is usually smaller than the number of replicate samples for the delete-1 jackknife method.

Concerns with survey confidentiality provide perhaps the sharpest distinction for you to choose among the different methods in SAS software. Some surveys release information to the public but also protect the respondent from being individually identifiable. In order to maintain this confidentiality, the strata or cluster identifications might not be released in the public use data sets. Instead the replicate weights are provided for variance estimation purposes. The Taylor series expansion method produces an incorrect estimate of variance if the strata or cluster information is unknown. However, the replicate weights contain all the necessary information for variance estimation, and hence replication variance estimation methods can be applied in these situations.

## OTHER ENHANCEMENTS IN SURVEY PROCEDURES

Other significant enhancements in the SAS 9.2 survey procedures include:

- domain estimation (SURVEYLOGISTIC and SURVEYREG)
- allocation of the total sample size among strata (SURVEYSELECT)
- estimation of the odds ratio and relative risks (SURVEYFREQ)
- finite population quantile estimation (SURVEYMEANS)

This section gives a brief description of these enhancements.

### DOMAIN ESTIMATION

Survey practitioners often compute statistics for a particular subpopulation or domain. For example, you might want to estimate the mean total health care expenditure for each poverty category for the 1999 population. Unless the survey is designed specifically to control the sample size in each domain, the randomness of domain sizes needs to be taken into account, in effect using the entire sample to estimate variance for domain estimates. In general, a subset analysis that uses the **BY** or **WHERE** statement provides incorrect estimate of variance for domain estimates. For more information about domain analysis, see Lohr (1999) and Särndal, Swenson, and Wretman (1992). All survey analysis procedures in SAS/STAT software—SURVEYMEANS, SURVEYFREQ, SURVEYREG, and SURVEYLOGISTIC—now implement domain estimation techniques. You can use the **DOMAIN** statement in the SURVEYMEANS, SURVEYREG or SURVEYLOGISTIC procedures to request analysis for domains in addition to analysis for the entire study population. To request domain analysis with PROC SURVEYFREQ, include one or more domain variables in your **TABLES** statement request.

### ALLOCATION OF THE TOTAL SAMPLE SIZE AMONG STRATA

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. Suppose you want to select a stratified random sample from a finite population. It is reasonable to allocate the total sample size among the strata based on stratum sizes, stratum variances, or stratum costs. The stratum cost is defined as the cost of observing one unit in the stratum. The SURVEYSELECT procedure in SAS/STAT 9.2 can allocate the total sample size among the strata based on a specified allocation method. You can request proportional allocation, Neyman allocation, or optimal allocation. Proportional allocation assigns the total sample size among the strata in proportion to the number of sampling units in the strata. The Neyman allocation assigns the total sample size among the strata in proportion to the strata sizes and strata variances. For Neyman allocation, you need to specify the stratum variance for each stratum. Optimal allocation assigns the total sample size among the strata in proportion to strata sizes, strata costs, and strata variances. For optimal allocation, you need to specify stratum variance and stratum cost for each stratum. You can use the **ALLOC=** option in the **STRATA** statement to request allocation of the total sample size among the strata. You can specify the total sample size in the **SAMPLESIZE =** option in the **PROC SURVEYSELECT** statement. **ALLOC=PROP** requests a proportional allocation, **ALLOC=NEYMAN** requests the Neyman allocation, and **ALLOC=OPTIMAL** requests an optimal allocation. See Chapter 87, “The SURVEYSELECT Procedure” (*SAS/STAT User’s Guide*), for details.

### ESTIMATION OF THE ODDS RATIO AND RELATIVE RISKS

The SURVEYFREQ procedure in SAS 9.2 also provides estimates for the odds ratio, relative risks, column risks, and risk differences for 2×2 tables. For the 1999 MEPS population, the odds of being uninsured for males is defined as the ratio of the number of males without any health insurance in 1999 and the number of males with some health insurance

in 1999. Similarly, the odds of being uninsured for females is defined as the ratio of the number of females without any health insurance in 1999 and the number of females with some health insurance in 1999. Suppose you want to estimate the ratio of the odds of being uninsured for males and the odds of being uninsured for females for the 1999 population. You can use the `OR` option in the `TABLE` statement to request the odds ratio. The following SAS statements estimate the odds ratio and relative risks for the 1999 population.

```
proc surveyfreq data = exampledata varmethod = taylor;
  strata varianceStrata;
  cluster variancePSU;
  weight personWeight;
  tables sex*insured / or;
run;
```

You can specify `VARMETHOD = TAYLOR`, `VARMETHOD = JACKKNIFE` or `VARMETHOD = BRR` to request different variance estimation methods. You can use the `RISK` option in the `TABLE` statement to request column risks and risk differences. See Chapter 83, "The SURVEYFREQ Procedure" (*SAS/STAT User's Guide*), for details.

Figure 10 displays the estimated odds ratio and relative risk produced by PROC SURVEYFREQ. The odds ratio is estimated as 1.31 and the 95% confidence interval is (1.21, 1.43). This result shows that the odds of being uninsured for males is higher than the odds of being uninsured for the females for the 1999 MEPS population.

Figure 10 Odds Ratio and Relative Risks

The SURVEYFREQ Procedure			
Table of SEX by insured			
Odds Ratio and Relative Risks (Row1/Row2)			
	Estimate	95% Confidence Limits	
Odds Ratio	1.3128	1.2061	1.4288
Column 1 Relative Risk	1.2744	1.1814	1.3747
Column 2 Relative Risk	0.9707	0.9619	0.9797
Sample Size = 23565			

## FINITE POPULATION QUANTILE ESTIMATION

Often survey data are used to estimate finite population quantiles, most commonly the median. You can use the SURVEYMEANS procedure in SAS/STAT 9.2 to estimate finite population quantiles and their standard errors. For example, the following SAS statements estimate the first quartile, median, and the third quartile for total health care expenditure for an individual for the 1999 population.

```
proc surveymeans data = exampledata quantile=(0.25,0.5,0.75);
  strata varianceStrata;
  cluster variancePSU;
  weight personWeight;
  var expenditure;
run;
```

The PROC statement option `QUANTILE = (0.25, 0.5, 0.75)` requests estimations of the first quartile, median, and the third quartile, respectively, for the analysis variable. You can also use the `PERCENTILE=(values)` option in the PROC SURVEYMEANS statement to request arbitrary percentiles. PROC SURVEYMEANS uses the Woodruff method (Särndal, Swenson, and Wretman 1992; Francisco and Fuller 1991) to estimate the variances of the estimated quantiles. See Chapter 85, "The SURVEYMEANS Procedure" (*SAS/STAT User's Guide*), for details.

Figure 11 displays the estimated quantiles for health care expenditure and their standard errors produced by PROC SURVEYMEANS. The median health care expenditure is estimated as 451.06 with a standard error of 12.55. A 95% confidence interval for the population median is (426.36, 475.75). It is not surprising that the estimated median health care expenditure (451.06) is much lower than the estimated mean health care expenditure (2156.47) because a large portion of the 1999 population has a zero or low health care expenditure.

**Figure 11** Quantiles for Health Care Expenditure

The SURVEYMEANS Procedure										
Quantiles										
Variable	Label	Percentile	Estimate	Std Error	95% Confidence Limits					
expenditure	TOTAL HEALTH CARE EXP 99	25% Q1	92.335200	3.930180	84.60267	100.06773				
	TOTAL HEALTH CARE EXP 99	50% Median	451.056590	12.550620	426.36355	475.74963				
	TOTAL HEALTH CARE EXP 99	75% Q3	1640.360608	41.935227	1557.85407	1722.86715				

## MISSING VALUES

Missing values in your survey data can compromise the quality of your survey results. They can arise for many reasons, such as coding errors or nonresponse. If the respondents are different from the nonrespondents with regard to a survey effect or outcome, then survey estimates might be biased and cannot accurately represent the survey population. There are a variety of techniques in sample design and survey operations that can reduce nonresponse. After data collection is complete, you can use imputation to replace missing values with acceptable values, or you can use sampling weight adjustments to compensate for nonresponse, or you can use both. You should complete this data preparation and adjustment before you analyze your data. See Kish (1965), Cochran (1977), Kalton (1983), and Brick and Kalton (1996) for more information. If an observation has a missing value or a nonpositive value for the **WEIGHT** variable, then that observation is excluded from the analysis. An observation is also excluded if it has a missing value for design variables such as **STRATA** variables, **CLUSTER** variables, and **DOMAIN** variables, unless missing values are regarded as a legitimate categorical level for these variables, as specified by the **MISSING** option. In addition to these exclusion conditions, the **SURVEYMEANS**, **SURVEYREG**, and **SURVEYLOGISTIC** procedures also omit observations with missing values for any analysis variable; these procedures compute statistics based only on observations that have nonmissing values, in effect treating the nonrespondents as *missing completely at random* (MCAR).

However, the number of nonmissing observations is often random. The **NOMCAR** option in the **PROC** statement takes into account this variability when estimating variance for an estimator. When you specify the **NOMCAR** option, the procedure treats observations with and without missing values for analysis variables as two different domains, and it performs a domain analysis in the domain of nonmissing observations. When the replicate weights are generated from the full sample by using both the missing and nonmissing observations, the replication methods automatically perform a domain analysis in the set of nonmissing observations. Hence the **NOMCAR** option is not required for replication methods.

## CONCLUSION

The three most commonly used and well-studied variance estimation methods for survey data are available in SAS 9.2—namely, Taylor series expansion, jackknife, and balanced repeated replication. Flexible implementation of replication variance estimation methods in SAS can be used for a wide variety of estimators. Other significant enhancements such as optimal allocations, domain estimations, quantile estimations, and estimations of the odds ratio are valuable additions for design and analysis of survey data.

## APPENDIX: BRR FOR DESIGNS WITH MORE THAN TWO PSUs PER STRATUM

Suppose your design has more than two PSUs in some or all strata. Two simple methods to implement balanced half sample techniques for such designs are the grouped balanced half samples (GBHS) (Kish and Frankel 1968; Wolter 1985; Shao and Tu 1995) and the repeated grouped balanced half samples (RGBHS) (Rao and Shao 1996; Shao and Tu 1995). The GBHS method randomly assigns PSUs into two groups of approximately the same size in each stratum. The BRR method is then applied to the two groups in each stratum instead of to individual PSUs. You can use a SAS data step to redefine PSU identifications in order to implement the GBHS method.



The following SAS statements use the 1999 MEPS data and randomly divide the original PSUs in each stratum into two groups. The data set `brrexample` contains all the variables from the data set `exempladata`, in addition to the group identification variable named `brbpsu`. Note that the random grouping of PSUs are not required to implement the BRR method for a two PSU per stratum design. If you have a two PSU per stratum design then you do not need to use these SAS statements.

```
proc freq data = exempladata noprint;
  table varianceStrata*variancePSU / out = freqstrpsu;
run;

data freqstrpsu; set freqstrpsu;
  rand = ranuni(2211);
proc sort data=freqstrpsu;
  by varianceStrata rand;
data brrstrpsu; set freqstrpsu;
  brrbpsu = mod(_N_,2);
run;

proc sort data = exempladata;
  by varianceStrata variancePSU;
proc sort data = brrstrpsu;
  by varianceStrata variancePSU;
run;
data brrexample;
  merge exempladata brrstrpsu (keep = varianceStrata variancePSU brrbpsu);
  by varianceStrata variancePSU;
run;
```

## REFERENCES

- Berger, Y. G. and Skinner, C. J. (2005), "A Jackknife Variance Estimator for Unequal Probability Sampling," *Journal of the Royal Statistical Society, Series B*, 67(1), 79–89.
- Brick, J. M. and Kalton, G. (1996), "Handling Missing Data in Survey Research," *Statistical Methods in Medical Research*, 5, 215–238.
- Chamber, R. L. and Skinner, C. J. (2003), *Analysis of Survey Data*, Chichester: John Wiley & Sons.
- Cochran, W. G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley & Sons.
- Dippo, C. S., Fay, R. E., and Morganstein, D. H. (1984), "Computing Variances from Complex Samples with Replicate Weights," in *Proceedings of the Survey Research Methods Section*, 489–494, American Statistical Association.
- Fay, R. E. (1989), "Theory and Application of Replicate Weighting for Variance Calculations," in *Proceedings of the Survey Research Methods Section*, 212–217, American Statistical Association.
- Francisco, C. A. and Fuller, W. A. (1991), "Quantile Estimation with a Complex Survey Design," *Annals of Statistics*, 19, 454–469.
- Fuller, W. A. (2006), "Sampling Statistics," Unpublished manuscript.
- Judkins, D. R. (1990), "Fay's Method for Variance Estimation," *Journal of Official Statistics*, 6(3), 223–239.
- Kalton, G. (1983), *Introduction to Survey Sampling*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-035, Beverly Hills and London: Sage Publications.
- Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons.
- Kish, L. and Frankel, M. R. (1968), "Balanced Repeated Replication for Analytical Statistics," in *Proceedings of the Social Statistics Section*, 2–11, American Statistical Association.
- Kish, L. and Frankel, M. R. (1970), "Balanced Repeated Replication for Standard Errors," *Journal of the American Statistical Association*, 65(331), 1071–1094.
- Kish, L. and Frankel, M. R. (1974), "Inference from Complex Samples," *Journal of the Royal Statistical Society, Series B*, 36(1), 1–37.
- Kovar, J. G., Rao, J. N. K., and Wu, C. F. J. (1988), "Bootstrap and Other Methods to Measure Errors in Survey Estimates," *The Canadian Journal of Statistics*, 16, 25–44.

- Krewski, D. and Rao, J. N. K. (1981), "Inference from Stratified Samples: Properties of the Linearization, Jackknife, and Balanced Repeated Replication Methods," *Annals of Statistics*, 9(5), 1010–1019.
- Lohr, S. L. (1999), *Sampling: Design and Analysis*, Pacific Grove, CA: Duxbury Press.
- Rao, J. N. K. and Shao, J. (1996), "On Balanced Half-Sample Variance Estimation in Stratified Random Sampling," *Journal of the American Statistical Association*, 91(433), 343–348.
- Rao, J. N. K. and Shao, J. (1999), "Modified Balanced Repeated Replication for Complex Survey Data," *Biometrika*, 86(2), 403–415.
- Rao, J. N. K. and Wu, C. F. J. (1983), *Inference from Stratified Samples: Second Order Analysis of Three Methods for Nonlinear Statistics*, Technical Report 7, Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, Ottawa.
- Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics*, 1(4), 381–397.
- Särndal, C. E., Swenson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag Inc.
- Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer-Verlag.
- Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.

## CONTACT INFORMATION

Pushpal K Mukhopadhyay  
SAS Institute Inc.  
100 SAS Campus Drive  
Cary, NC, 27513  
Work Phone: 919-531-2123  
E-mail: pushpal.mukhopadhyay@sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.