Paper 267-2010

Using the SAS® Survey Procedures for Subpopulation Analysis with Jackknife Repeated Replication Methods in SAS® 9.2

Zhuoqiao Wang, Information Management Services, Inc., Silver Spring, MD
William Waldron, Information Management Services, Inc., Silver Spring, MD

ABSTRACT

In the analysis of complex sample data, Jackknife Repeated Replication is a method to estimate the variance of a statistic that can result in substantially reduced variance when compared to traditional estimators. Subpopulation analysis is used to compute point and variance estimates for a subset of the sampled population. This is accomplished through a judicious adjustment of the final sample weights. SAS® 9.2 supports the Jackknife Repeated Replication method. However, the subpopulation analysis for Jackknife Repeated Replication is not available in SAS® 9.2. The DOMAIN statement is available only for the Taylor series method. This paper presents how to conduct a general subpopulation analysis for jackknife replicate weight designs using the Health Information National Trends Survey (HINTS) 2005 data. Mean estimation and logistic regression are demonstrated. The examples are illustrated with SAS® 9.2, and the results are compared to standard subpopulation analyses from Stata 10.0 SE and SUDAAN 10.

INTRODUCTION

Jackknife Repeated Replication (JRR, delete-one jackknife in this paper) is a commonly used resampling approach to variance estimation. In survey data containing H strata and n_h Primary Sampling Units (PSUs) in stratum h, the replicate weight for the subsample after deleting one PSU from stratum h is,

$$w^*_{a,b,j} = \begin{cases} 0 & \text{if } a = h \text{ and } b \text{ is dropped} \\ \dfrac{n_h}{n_h - 1} w_{a,b,j} & \text{if } a = h \text{ and } b \text{ is not dropped} \\ w_{a,b,j} & \text{otherwise} \end{cases}$$

where $w_{a,b,j}$ is the full sample weight for the jth individual from bth PSU in stratum a.

For a full sample parameter vector θ (e.g., means, totals, ratios or regression coefficients), the $\hat{\theta}_{(h,i)}(w^*_{a,b,j})$ is a vector of point estimates for the replicate after deleting the ith PSU from stratum h and the $\hat{\theta}(w_{a,b,j})$ is a vector of point estimates for the full sample.

The jackknife mean is

$$\bar{\theta}_{(.)} = \frac{1}{n}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\{\hat{\theta}_{(h,i)}\,(w^*_{a,b,j})\}$$

where $n = \sum_{h=1}^{L} n_h$is the total number of PSUs in the full sample.

The variance estimator is

$$\hat{V}(\hat{\theta}) = \sum_{h=1}^{L} m_h \sum_{i=1}^{n_h}\{\hat{\theta}_{(h,i)}(w^*_{a,b,j}) - \hat{\theta}(w_{a,b,j})\}\{\hat{\theta}_{(h,i)}(w^*_{a,b,j}) - \hat{\theta}(w_{a,b,j})\}'$$

where $m_h = \frac{n_h - 1}{n_h}$ is the multiplier associated with stratum h for delete-one jackknife.

Once the replicates are created, the information regarding strata and PSUs is not necessary and will be removed in practice. The total number of replicates is denoted as $R = n = \sum_{h=1}^{L} n_h$.

The jackknife mean is rewritten as

$$\bar{\theta}_{(.)} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r(w_r)$$

where w_r is associated with $w^*_{a,b,j}$ and $\hat{\theta}_r(w_r)$ is associated with $\hat{\theta}_{(h,i)}(w^*_{a,b,j})$.

The variance estimator is rewritten as

$$\hat{V}(\hat{\theta}) = \sum_{r=1}^{R} m_r\{\hat{\theta}_r(w_r) - \hat{\theta}(w)\}\{\hat{\theta}_r(w_r) - \hat{\theta}(w)\}'$$

where w is associated with $w_{a,b,j}$, w_r is associated with $w^*_{a,b,j}$, $\hat{\theta}_r(w_r)$ is associated with $\hat{\theta}_{(h,i)}(w^*_{a,b,j})$, $\hat{\theta}(w)$ is associated with $\hat{\theta}(w_{a,b,j})$ and m_r is associated with m_h.

Subpopulation analysis is appropriate if the interest focuses on part of the sampled population. For instance, a researcher may only be concerned with a particular gender or age group within the sample population. For the point estimate in a subpopulation analysis, the sample weights can be set to zero for sampled individuals outside the subpopulation while the sample weights are kept unchanged for sampled individuals within the subpopulation. Weighted estimates can be calculated using the entire sampled data with the newly adjusted sample weights. In the subpopulation analysis of JRR, both the sample weights and the replicate weights should be adjusted based on the subpopulation of interest. Please see Cochran (1977), Graubard and Korn (1996), Korn and Graubard (1999), Oh and Scheuren (1983), Research Triangle Institute (2008), StataCorp (2007) and Wolter (1985) for details.

Though the JRR for subpopulation analysis is a standard feature in Stata 10.0 and SUDAAN 10, it is not yet available in SAS® 9.2. Thus, this paper demonstrates examples of subpopulation analysis with jackknife replicate weights using a reweighting method: adjust the weights in DATA step; use the newly adjusted weights in PROC SURVEYMEANS procedure and PROC SURVEYLOGISTIC procedure for subpopulation analysis. The results from SAS® 9.2 are compared to the standard subpopulation analyses from Stata 10.0 and SUDAAN 10. The choice of convergence criteria in PROC SURVEYLOGISTIC and the construction of 95% confidence interval for point estimate are discussed.

METHOD

In a subpopulation analysis for survey data with the JRR method, the indicator variable I_D is created for the j$^{\text{th}}$ individual in the i$^{\text{th}}$ PSU in the h$^{\text{th}}$ stratum for a domain D.

$$I_D(h,i,j) = \begin{cases} 1 & \text{if observation } (h,i,j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

The indicator variable I_D will be applied to the full sample weights because the full sample weights are used to calculate the $\hat{\theta}(w_{a,b,j})$ on the domain D. Similarly, the indicator variable I_D will be applied to all replicate weights because the replicate weights are used to calculate the corresponding $\hat{\theta}_{(h,i)}(w^*_{a,b,j})$ on the domain D.

The new full sample weights for domain D are

$$v_{a,b,j} = w_{a,b,j} I_D(a,b,j) = \begin{cases} w_{a,b,j} & \text{if observation (a, b, j)belongs to D} \\ 0 & \text{otherwise} \end{cases}$$

The new replicate weights for domain D are

$$v^*_{a,b,j} = w^*_{a,b,j} I_D(a,b,j) = \begin{cases} w^*_{a,b,j} & \text{if observation } (a,b,j)\text{belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

In the above formula, the new full sample weights and new replicate weights are created by applying the indicator variable I_D to the original full sample weights and the original replicate weights, respectively. Then, the point estimate for the parameter of interest on the domain D is computed by using the new replicate weights. The variance of the statistic of interest on the domain D is estimated by the variability of the point estimates using the new replicate weights deviated from the point estimate using the new full sample weights.

Without the notation of strata and PSUs, the jackknife mean on the domain D is,

$$\bar{\theta}_{(.)} = \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}_r(v_r)$$

where v_r is associated with $v^*_{a,b,j}$.

The variance estimator on the domain D is,

$$\hat{V}(\hat{\theta}) = \sum_{r=1}^{R} m_r\{\hat{\theta}_r(v_r) - \hat{\theta}(v)\}\{\hat{\theta}_r(v_r) - \hat{\theta}(v)\}'$$

where v_r is associated with $v^*_{a,b,j}$ and v is associated with $v_{a,b,j}$. Please see details in Graubard and Korn (1996), Korn and Graubard (1999), Lumley (2004) and StataCorp (2007).

This paper demonstrates how to construct the new full sample weights and the new replicate weights for a subpopulation of interest using SAS® DATA step and compute the point and variance estimates for the corresponding subpopulation by using the new full sample weights and the new replicate weights in survey procedures in SAS® 9.2.

HINTS 2005 DATA

The Health Information National Trends Survey (HINTS) is a national, biennial survey designed to collect nationally representative data about the American public's use of cancer-related information. This survey is a leading source of data on health communication issues and is sponsored and developed by National Cancer Institute's Division of Cancer Control and Population Sciences. The baseline year is 2003 while HINTS 2007 represents the third round data of collection. The examples in this paper use HINTS 2005 data.

The HINTS 2005 data (second round) were collected from February 2005 through August 2005. The sample is drawn, using Random Digit Dial (RDD) telephone numbers via a Computer Assisted Telephone Interview (CATI) format, to produce a nationally representative sample of telephone households. During the household screening, one adult was sampled within each household and recruited for the extended interview.

In HINTS 2005, the replicate weights are based on a delete-one jackknife method. Each sampled telephone number was assigned to one of 50 replicate "deletion" groups. Each replicate sample is the full sample minus the deletion group and the corresponding replicate weight is formed by reweighting the replicate sample as if it was the full sample. Thus, there are 50 replicate weights and the multiplier is 49/50=0.98. More information regarding HINTS is available at http://hints.cancer.gov.

The details of the analyses variables and demographic variables used in this paper can be found in Rizzo et al (2008). They are:

InternetForCancer: Have you ever specifically looked for cancer information online?

Age: Age category variable

Sex: Gender variable

Educ: Education category variable

The weight variables are:

fwgt: the original full sample weights

fwgt1-fwgt50: the original replicate weights

The education category variable (educ) will be used in the following illustrations of subpopulation analysis.

The following SAS® statements subset the combined data set (2003 and 2005) to get the HINTS 2005 data. The new weights (nfwgt, nfwgt1-nfwgt100) for the trend analysis were dropped.

```
data hints2005;
    set combined;

    drop nfwgt nfwgt1-nfwgt100;
    format InternetForCancer srvyYear educ age sex;

    where srvyYear = 2;
run;
```

JACKKNIFE REPEATED REPLICATION VARIANCE ESTIMATION IN SAS® 9.2

The following SAS® statements are used to demonstrate the subpopulation analysis using the JRR variance estimator. Neither the results nor the code should be used for inferential purposes.

SYNTAX

The option VARMETHOD=JACKKNIFE | JK <(method-options)> in the corresponding survey PROC statement will request the Jackknife Repeated Replication variance estimation method.

The WEIGHT statement names the variable that contains the sampling weights, which is the full sampling weights. The sampling weights must be positive numbers.

The statement REPWEIGHTS variables < / options>, names the variables that provide replicate weights for BRR or jackknife variance estimation. The replicate weights must be nonnegative numbers (zero or positive). The option JKCOEFS=value will be used if the multipliers (jackknife coefficient in SAS®) m_r are the same across all replicates. The option JKCOEFS=values or JKCOEFS=SAS-data-set should be used if the multipliers m_r are different.

The statement DOMAIN designed for subpopulation analysis is available only for VARMETHOD=TAYLOR, which motivates us to conduct subpopulation analysis using the method below.

ESTIMATE THE MEANS OR PROPORTIONS

The InternetForCancer variable is binary: 0 if a person had not looked for health information online; 1 if a person had looked for health information online. The mean of InternetForCancer is the proportion of people who had looked for health information online. The variable Educ has four levels: 1 (less than high school graduate), 2 (high school graduate), 3 (some college) and 4 (college graduate). The following codes demonstrate how to create the indicator variable for the specified subgroup, create the new weight variables, and estimate the mean of a numeric variable (or a binary variable) for the specified subgroup.

```
%macro subpop(DSIN=, SWGT=, PRERWGT=, NUMRWGT=, MULTIPLIER=, SUBPOPVAR=,
SUBPOPLEVEL=, VAR=);
    data &DSIN.copy;
           set &DSIN.;
           array a(*) &PRERWGT.1-&PRERWGT.&NUMRWGT.;

* Create the indicator variable for the specified value(s) of subpopulation
variable;
* Delete the individuals having missing value of subpopulation variable;
           if missing(&SUBPOPVAR.) = 0 then
                  do;
                         if &SUBPOPVAR. = &SUBPOPLEVEL. then
                                _flag = 1;
                         else
                                _flag = 0;
                  end;
           else
                  delete;

* The new full sample weight variable are same as before if _flag = 1;
* The new full sample weight variable are assigned with 0.000001 if _flag = 0;
           if _flag = 0 and &SWGT. not in (., 0) then
                  &SWGT. = 0.000001;

* Apply the indicator variable to the original replicate weights;
           do _I = 1 to dim(a);
                  a(_I) = a(_I)*_flag;
           end;
    run;

* Conduct the analysis using the new weight variables.;
    proc surveymeans data=&DSIN.copy varmethod=jackknife nobs nmiss df mean stderr;
           weight &SWGT.;
           repweights &PRERWGT.1-&PRERWGT.&NUMRWGT. / jkcoefs=&MULTIPLIER.;
           var &VAR.;
           ods output Statistics=OutStatistics;
    run;

    proc print data=OutStatistics noobs;
           format mean stderr 12.7;
    run;
%mend subpop;

%subpop(DSIN=hints2005, SWGT=fwgt, PRERWGT=fwgt, NUMRWGT=50, MULTIPLIER=0.98,
SUBPOPVAR=educ, SUBPOPLEVEL=1, VAR=InternetForCancer);

%subpop(DSIN=hints2005, SWGT=fwgt, PRERWGT=fwgt, NUMRWGT=50, MULTIPLIER=0.98,
SUBPOPVAR=educ, SUBPOPLEVEL=2, VAR=InternetForCancer);

%subpop(DSIN=hints2005, SWGT=fwgt, PRERWGT=fwgt, NUMRWGT=50, MULTIPLIER=0.98,
SUBPOPVAR=educ, SUBPOPLEVEL=3, VAR=InternetForCancer);

%subpop(DSIN=hints2005, SWGT=fwgt, PRERWGT=fwgt, NUMRWGT=50, MULTIPLIER=0.98,
SUBPOPVAR=educ, SUBPOPLEVEL=4, VAR=InternetForCancer);
```

The statistic keywords MEAN and STDERR request the mean and its standard error respectively, for the variable of interest specified in the statement VAR.

In the above code, the individuals having missing value in subpopulation variable are deleted from the data set for subpopulation analysis per definition. A very small, positive full sample weight (instead of a zero full sample weight) is assigned to the individuals falling outside the subpopulation because SAS® omits the observations with missing or non-positive full sample weight. Please see the discussion in Graubard and Korn (1996).

ESTIMATE THE LOGISTIC REGRESSION COEFFICIENTS

The logistic regression model can be used to examine the association between usage of internet for cancer information and demographic variables. The following SAS® statements demonstrate how to create the indicator variable for the specified subgroup, create the new weight variables and estimate the logistic regression coefficients for the specified subgroup.

```
%macro subpop2(DSIN=, SWGT=, PRERWGT=, NUMRWGT=, MULTIPLIER=, SUBPOPVAR=,
SUBPOPLEVEL=);
    data &DSIN.copy;
           set &DSIN.;
           array a(*) &PRERWGT.1-&PRERWGT.&NUMRWGT.;

* Create the indicator variable for the specified value(s) of subpopulation
variable;
* Delete the individuals having missing value of subpopulation variable;
           if missing(&SUBPOPVAR.) = 0 then
                  do;
                         if &SUBPOPVAR. = &SUBPOPLEVEL. then
                                _flag = 1;
                         else
                                _flag = 0;
                  end;
           else
                  delete;

* The new full sample weight variable are same as before if _flag = 1;
* The new full sample weight variable are assigned with 0.000001 if _flag = 0;
           if _flag = 0 and &SWGT. not in (., 0) then
                  &SWGT. = 0.000001;

* Apply the indicator variable to the original replicate weights;
           do _I = 1 to dim(a);
                  a(_I) = a(_I)*_flag;
           end;
    run;

    proc surveylogistic data=&DSIN.copy varmethod=jackknife;
           weight &SWGT.;
           class age (ref='4') sex (ref='2') / param=ref;
           repweights &PRERWGT.1-&PRERWGT.&NUMRWGT. / jkcoefs=&MULTIPLIER.;
           model InternetForCancer (descending) = age sex / tech=newton xconv=1E-6;
           ods output parameterestimates=pe;
    run;

    proc print data=pe noobs;
           format Estimate StdErr 12.7;
    run;
%mend subpop2;
```

```
%subpop2(DSIN=hints2005, SWGT=fwgt, PRERWGT=fwgt, NUMRWGT=50, MULTIPLIER=0.98,
SUBPOPVAR=educ, SUBPOPLEVEL=1);

%subpop2(DSIN=hints2005, SWGT=fwgt, PRERWGT=fwgt, NUMRWGT=50, MULTIPLIER=0.98,
SUBPOPVAR=educ, SUBPOPLEVEL=2);

%subpop2(DSIN=hints2005, SWGT=fwgt, PRERWGT=fwgt, NUMRWGT=50, MULTIPLIER=0.98,
SUBPOPVAR=educ, SUBPOPLEVEL=3);

%subpop2(DSIN=hints2005, SWGT=fwgt, PRERWGT=fwgt, NUMRWGT=50, MULTIPLIER=0.98,
SUBPOPVAR=educ, SUBPOPLEVEL=4);
```

The statement CLASS age (ref='4') sex (ref='2') / PARAM=ref, requests 65 or older (age=4) as the referent group for age category variable and female (sex=2) as the referent group for the gender variable, respectively.

The option DESCENDING for the variable InternetForCancer in the MODEL statement indicates that having looked for cancer information online (InternetForCancer=1) is the event of interest.

The option TECH=newton in the MODEL statement requests the Newton-Raphson algorithm as the optimization technique during the model fitting to replace the default Fisher scoring algorithm. The Stata 10.0 uses a modified Newton-Raphson algorithm as default. The SUDAAN 10 has only a modified Newton-Raphson algorithm for the logistic regression. Thus, the option TECH=newton allows us to compare the estimate and standard error from SAS® 9.2 to those from Stata 10.0 and SUDAAN 10.

The option XCONV=1e-6 in MODEL statement requests the relative parameter change as the convergence criterion to replace the default relative gradient convergence criterion and specifies the value. The Stata 10.0 uses double convergence criteria: a scaled gradient, and either relative change in the parameter or relative change in log likelihood. The SUDAAN 10 uses the relative change in parameter at default. The option XCONV=1e-6 is suitable for the comparisons among different software packages.

Again, the individuals having missing value in subpopulation variable are deleted from the data set for subpopulation analysis per definition. A very small, positive full sample weight (instead of a zero full sample weight) is assigned to the individuals falling outside the subpopulation.

RESULTS COMPARISON

Education subgroup	Mean (standard error) from SAS® 9.2	Mean (standard error) from Stata 10.0	Mean (standard error) from SUDAAN 10
Less than high school graduate	0.0640224 (0.0105911)	0.0640224 (0.0105911)	0.0640224 (0.0105911)
High school graduate	0.1994112 (0.0156818)	0.1994112 (0.0156818)	0.1994112 (0.0156818)
Some college	0.3467236 (0.0194107)	0.3467236 (0.0194107)	0.3467236 (0.0194107)
College graduate	0.4646702 (0.0161146)	0.4646702 (0.0161146)	0.4646702 (0.0161146)

Table 1. Mean and Standard Error of InternetForCancer for Education Subgroup

The results of each education subgroup from PROC SURVEYMEANS in SAS® 9.2 using the modified weights are almost same as those from the standard subpopulation analysis procedure in Stata 10.0 and SUDAAN 10.

Education Subgroup	Parameter	Estimate (standard error) from SAS® 9.2	Estimate (standard error) from Stata 10.0	Estimate (standard error) from SUDAAN 10
Less than high school graduate	intercept	-3.4884920 (0.4637221)	-3.488492 (0.4637222)	-3.488492 (0.4637221)
	18-34 years old	1.4545255 (0.5450948)	1.454526 (0.5450948)	1.4545255 (0.5450948)
	35-49 years old	1.3144992 (0.5645722)	1.314499 (0.5645722)	1.3144992 (0.5645722)
	50-64 years old	0.5435616 (0.7327003)	0.5435616 (0.7327003)	0.5435616 (0.7327003)
	female	-0.4623270 (0.4524979)	-0.462327 (0.4524979)	-0.462327 (0.4524979)
High school graduate	intercept	-2.6355384 (0.2524743)	-2.635538 (0.2524742)	-2.635538 (0.2524743)
	18-34 years old	1.7534778 (0.2885113)	1.753478 (0.2885114)	1.7534778 (0.2885113)
	35-49 years old	1.6340474 (0.2634412)	1.634047 (0.2634411)	1.6340474 (0.2634412)
	50-64 years old	1.4052223 (0.3259096)	1.405222 (0.3259096)	1.4052223 (0.3259096)
	female	-0.3833062 (0.2665393)	-0.3833062 (0.2665393)	-0.383306 (0.2665393)
Some college	intercept	-1.8204201 (0.1934062)	-1.82042 (0.1934062)	-1.820420 (0.1934062)
	18-34 years old	1.5261572 (0.2296885)	1.526157 (0.2296884)	1.5261572 (0.2296885)
	35-49 years old	1.6035376 (0.2272878)	1.603538 (0.2272877)	1.6035376 (0.2272878)
	50-64 years old	1.3275663 (0.2365501)	1.327566 (0.2365501)	1.3275663 (0.2365501)
	female	-0.4267207 (0.1727984)	-0.4267207 (0.1727984)	-0.426721 (0.1727984)
College graduate	intercept	-0.9632558 (0.1666569)	-0.9632558 (0.1666569)	-0.963256 (0.1666569)
	18-34 years old	1.1912719 (0.2039414)	1.191272 (0.2039414)	1.1912719 (0.2039414)
	35-49 years old	0.8582038 (0.1751153)	0.8582038 (0.1751153)	0.8582038 (0.1751153)
	50-64 years old	1.0201150 (0.1920158)	1.020115 (0.1920158)	1.0201150 (0.1920158)
	female	-0.1408337 (0.1241795)	-0.1408337 (0.1241795)	-0.140834 (0.1241795)

Table 2. Coefficients and Standard Error from Logistic Regression for Education Subgroup

With the Newton-Raphson algorithm and relative parameter change as the convergence criterion, the results of each education subgroup from PROC SURVEYLOGISTIC in SAS® 9.2 using the modified weights are almost same as those subpopulation analyses from the corresponding standard procedure in Stata 10.0 and SUDAAN 10.

DISCUSSION

The construction of 95% confidence interval is not demonstrated above since it involves two open questions: the underlying distribution of the test statistic and its degrees of freedom if applicable. To begin with, statistical software packages use different strategies to compute the degrees of freedom. Both SAS® 9.2 and SUDAAN 10 use the number of replicates as the default degrees of freedom and provide the option for user-specified degrees of freedom.

Stata 10.0 uses the number of replicates minus one as the default degrees of freedom and does not provide the option for user-specified degrees of freedom. Secondly, for the JRR variance estimation, the deviation of a replicate estimate $\hat{\theta}_r$ from the full sample estimate $\hat{\theta}$ could be negligibly close to zero in the subpopulation analysis. Thus, those replicates should not contribute to the degrees of freedom per Rizzo et al (2008). Generally, it requires manual work to determine which replicate has negligible contribution to the variance using the interim results of replicate estimates. None of the mainstream software provides this capability. Finally, regarding the construction of 95% confidence intervals for the coefficients in the logistic regression, both Stata 10.0 and SUDAAN 10 assume the test statistic of the coefficient follows a t-distribution. But, SAS® 9.2 assumes the test statistic of the coefficient follows a Wald chi-square distribution. In practice, there should not be much difference in the 95% confidence interval if the number of replicates is large enough.

The convergence criterion should be cautiously chosen in survey analysis using SAS® 9.2 with JRR method. The default relative gradient convergence criterion (GCONV=1E-8) could be questionable. It is designed as:

$$\frac{g^{(i)\prime} I^{(i)} g^{(i)}}{|l^{(i)}| + 1E - 6} < value$$

where $l^{(i)}$ is the value of the log-likelihood function, $g^{(i)}$ is the gradient vector and $I^{(i)}$ is the (expected) information matrix. All of them are evaluated at the ith iteration. The default value is 1E-8. The value of denominator (absolute value of log-likelihood function) is usually a large number in survey analysis, which will lead the whole left hand side to a much smaller number than the default value 1E-8 on right side and cause the iteration to stop early. In the above examples, the same results could be reached with GCONV=1E-16.

For the JRR variance estimation, the subpopulation analysis can be analyzed simply as a subset since the variance (standard error) is calculated with the point estimates, which are the same in the subsetting method and reweighting method (Lumley 2004). It is not in general correct to conduct subpopulation analysis using a subset of the original data because of adverse effects on the variance computation relating to total overall sample size. Therefore, it is always a good practice to conduct subpopulation analysis with the reweighting method.

CONCLUSION

This paper provides the practical guidance on the subpopulation analysis with Jackknife Repeated Replication method using the survey procedures in SAS® 9.2. The analyses were demonstrated using the public data. The issues regarding degrees of freedom, distribution of test statistic and choice of convergence criterion were also discussed. The method can be extended to other replicate weight designs.

REFERENCES

Cochran, W. G. (1977). Sampling Techniques, 3rd ed. New York: John Wiley & Sons.

Graubard, B. I., and Korn, E. L. (1996). Survey Inference for Subpopulations, American Journal of Epidemiology Vol. 144 No. 1

Korn, E. L., and Graubard, B. I. (1999). Analysis of Health Surveys. New York: John Wiley & Sons.

Lumley, T. (2004). Analysis of Complex Survey Samples, Journal of Statistical Software Vol. 9, Issue 8.

Oh, H. L., and Scheuren, F. S. (1983). Weighting adjustments for unit nonresponse, in Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography (W. G. Madow, I. Olkin, and D. B. Rubin, eds.), New York: Academic Press.

Research Triangle Institute (2008). SUDAAN Language Manual, Release 10.0 Research Triangle Park, NC: Research Triangle Institute.

Rizzo, L., Moser, R.P., Waldron., W., Wang, Z., and Davis, W.W. (2008). Analytic Methods To Examine Changes Across Years Using HINTS 2003 & 2005 Data. NIH Pub No. 08-6435.

StataCorp. 2007. Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

Wolter, K. M. (1985) Introduction to Variance Estimation. New York, NY:Springer.

ACKNOWLEDGMENTS

We wish to thank Dr. Dennis Buckman and Joe Zou for their helpful suggestions. We would also like to thank Dr. Richard Moser for helping to bring to our attention the complications that can arise when conducting statistical analysis using different software packages.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Zhuoqiao Wang
Information Management Services, Inc.
12501 Prosperity Drive, Suite 200
Silver Spring, MD 20904
wangz@imsweb.com

William Waldron
Information Management Services, Inc.
12501 Prosperity Drive, Suite 200
Silver Spring, MD 20904
waldronw@imsweb.com