ESTIMATING AHS-NATIONAL VARIANCES WITH REPLICATE WEIGHTS

Introduction

The American Housing Survey – National Sample (AHS-N) is a probability sample of housing units. Several methods can be used to estimate sampling variance for complex sample designs like the design used in the AHS-N. One computationally efficient method for calculating sampling variance is the method of replication, where replicate weights are used to estimate the variance. For the first time, the Census Bureau is releasing a replicate weight file that can be used with the AHS-N public use data file so that data users can generate variance estimates themselves with ease.

The variance of any survey estimate based on a probability sample may be estimated by the method of replication. This method requires that the sample selection, the collection of data, and the estimation procedures be independently carried through (replicated) several times. Each time the sample is replicated, a different set of estimates is calculated. The dispersion of the resulting estimates then can be used to measure the variance of the full sample (reference [1], Appendix D).

However, we would not consider repeating any large survey, such as the American Housing Survey – National Survey (AHS-N), several times to obtain variance estimates. A practical alternative is to manipulate the full sample several times by applying different weighting factors to the sample units. The manipulation of the weights causes the sample data to represent a different number of housing units in each replicated sample. We then apply the estimation procedures (e.g., mean, median, sum, etc.) to these weighted random samples. We refer to these random samples as replicates. For the AHS-N, we used a total of 160 replicates to calculate the AHS-N variance estimates.

Revision History

This document and the associated 2009 replicate weight files were revised in August 2012. The following revisions were made:

- Reordered the replicate weight files so that replicates 0 and 160 were exchanged, i.e. replictate 0 (REPWGT0) now represents the full sample weight (wgt90geo) and replicates 1-160 (REPWGT1-REPWGT160) now represents each of the 160 replicates.
- Changed the variable names in the replicate weights to reflect a reordering of the weights.
- Changed the ASCII input file to a comma delimited file with variable names.
- Assigned values of ‘B’ to the weights for the Type C noninterviews on the ASCII file, instead of values of -9. NOTE: the Type C noninterviews were already assigned values of ‘B’ in the SAS data set.
- Generalized the document so that it is not just applicable to 2009.
- Changed the statement to convert ASCII file to SAS dataset by specifying lrecl=5000.

THE REPLICATE WEIGHTS SHOULD ONLY BE USED IN CREATING VARIANCES AND SHOULD NOT BE USED TO CREATE INDEPENDENT ESTIMATES.

The user should also note that the replicate weights are calculated using information from the sample. Therefore; the 2009 AHS-N replicate weights are applicable for use on only 2009 AHS-N data, and so on. There are currently no plans to calculate replicate weights for the metropolitan surveys.

Use of Replicate Estimates in Variance Calculations

Calculate variance estimates using Fay's Balanced Repeated Replication (BRR) method (reference [2]) for AHS-N estimates using the following formula:

$$\text{var}(\hat{\theta}_0) = \frac{4}{160}\sum_{i=1}^{160}(\hat{\theta}_i - \hat{\theta}_0)^2 \qquad (1)$$

where $\hat{\theta}_0$ is the weighted estimate of the statistic of interest; such as a total, median (reference [3]), mean, proportion, regression coefficient, or log-odds ratio, using the weight for the full sample and $\hat{\theta}_i$ are the replicate estimates of the same statistic using the replicate weights. See reference [2]. To calculate a standard error, the measure of dispersion when parameter estimates are calculated through repeated sampling from the population, obtain the square root of the variance estimate.

To ensure confidentiality of the data, some characteristics have either been bottomcoded or topcoded. This procedure places a lower (or upper) boundary on the published value for the variable in question. Therefore, some estimates calculated from the Public Use File may differ from the estimates provided in the AHS-N publication tables. If a specific estimate is needed, contact the American Housing Survey Branch of the Census Bureau at (301) 763-3235 or ahsn@census.gov.

Example for Using Replication to Estimate Variances

The following example illustrates how a statistic would be estimated, replicated, and combined to form the variance estimate. To simplify calculations, we are going to estimate the variance using *k*=4 replicates rather than the 160 replicates provided for the AHS.

Please note that in the following example, the weights in Replicate 1 equal the Full-Sample Weight. In practice, this will not necessarily be the case, as the calculation of replicate weights is driven by the selection of the Hadamard matrix. See the unabridged version of this document for more details.

Sampling Variance of a Total

The goal of this example is to estimate the total number of owner-occupied housing units in our population and its corresponding estimate of variance. Assume we have five sample cases with responses shown below when asked if they owned their house (tenure) during the time of interview.

Sample Case	Owned House?	Full-Sample Weight	Replicate Weights			
			Replicate 1	Replicate 2	Replicate 3	Replicate 4
Case #1	**YES**	**15.96**	**15.96**	**5.30**	**24.90**	**15.84**
Case #2	NO	24.47	24.47	46.06	22.46	7.29
Case #3	**YES**	**20.21**	**20.21**	**22.38**	**5.57**	**34.11**
Case #4	**YES**	**17.02**	**17.02**	**18.85**	**26.56**	**5.07**
Case #5	NO	22.34	22.34	7.42	20.51	37.70

Step 1: Calculate the full-sample weighted survey estimate.

Add the full-sample weights of the sample cases that responded "YES" to the tenure question. Therefore, the total number of owner-occupied housing units survey estimate is calculated as follows:

- Full-Sample Owner-Occupied Estimate = 15.96 + 20.21 + 17.02 = **53.19**

Step 2: Calculate the weighted survey estimate for each of the replicate samples.

The replicate survey estimates are as follows:

- Replicate 1 Owner-Occupied Estimate = 15.96 + 20.21 + 17.02 = **53.19**
- Replicate 2 Owner-Occupied Estimate = 5.30 + 22.38 + 18.85 = **46.53**
- Replicate 3 Owner-Occupied Estimate = 24.90 + 5.57 + 26.56 = **57.03**
- Replicate 4 Owner-Occupied Estimate = 15.84 + 34.11 + 5.07 = **55.02**

Step 3: Use these survey estimates in formula (1) to calculate the variance estimate for the total owner-occupied population.

The calculation of this variance estimate is as follows:

$$\mathrm{Var}(\hat{Y}_0) = \frac{4}{\mathrm{k}} \sum_{i=1}^{\mathrm{k}} (\hat{Y}_i - \hat{Y}_0)^2,$$

where $\hat{Y}_0$ is the total calculated using the full-sample weight,

$\hat{Y}_i$ is the total calculated using replicate *i*, and

k is the number of replicates.

$$= \frac{4}{4} \times [(53.19 - 53.19)^2 + (46.53 - 53.19)^2 + (57.03 - 53.19)^2 + (55.02 - 53.19)^2]$$

$$= 1 \times [0 + 44.3556 + 14.7456 + 3.3489] = 62.4501.$$

Thus Var(y_o) = **62.4501.**

Step 4: Take the square root of the variance to get the standard error.

Therefore, the survey estimate for total owner-occupied population in our example is **53.19** housing units. This survey estimate has an estimated variance of **62.4501**, or a standard error of **7.90** housing units.

Use these four steps to calculate standard errors for other statistics. Examples for means, medians, proportions, regression parameters, and log-odds ratios are provided in the unabridged version of this document.

Confidence Intervals and Significance Tests

Once the standard error is calculated, it can be used to create confidence intervals and perform significance tests. Use the estimate where the equation requires the standard error. For means, medians, totals, and regression coefficients, the degrees of freedom will equal the number of replicates (see reference [7]). For detailed examples of proportions and log-odds ratios (as well as the other statistics given here), refer to the unabridged version of this document .

AHS-N Replicate Weight File Description

The file AHSN_ASCII_REPWGT_yyyy.TXT, found on The Department of Housing and Urban Development's website at: huduser.org, contains the replicate weights and match key required to merge the replicate weight file to the public use survey data file (yyyy is the data collection year). This is a comma delimited ASCII file consisting of the number of records on the AHS Public Use File and a header row with variable names. The maximum record length is 2102. The match key is the variable CONTROL (a character variable with length 12). When reading this file into SAS, the following example will provide the replicate weight data set (Input-related items are highlighted in yellow):

```
filename ASCII_file 'C:\file_pathway\AHSN_ASCII_REPWGT_20XX.txt';

data repwgt_20XX;
length control $12.;
infile ASCII_file dlm=',' lrecl=5000 firstobs=2;
input control repwgt0-repwgt160;
missing b;
run;
```

This SAS code identifies the variables in the data set, as well as the order in which they appear in the ASCII file. The first variable is CONTROL, the second variable is REPWGT0, the third is REPWGT1, and so on through REPWGT160. The replicate weights are in REPWGT1 – REPWGT160. The full sample weight is in REPWGT0.

The ASCII file and the public use survey data file both have the full sample weight. On the ASCII file the variable name is REPWGT0, but on the public use survey data file the variable name is WGT90GEO. The full weight on the ASCII file is given as means of verifying that the files are properly merged to the public use survey data.

Merging the AHS-N Replicate Weight File with the AHS-N Public Use File

Obtain: AHS-N Public Use File (file name = AHS20XX)
AHS-N Replicate Weight File (file name = REPWGT_20XX, or use ASCII file)

Merge using CONTROL. Before merging, ensure each file is sorted by CONTROL.

Examples of Calculating Variances Using Statistical Software

As of August 2012, documentation that describes variance estimation using Fay's BRR method can be accessed through the internet for the following applications:

R
Documentation for package 'survey' version 3.6-12, User Guides and Package Vignettes (http://rss.acs.unt.edu/Rdoc/library/survey/html/00Index.html)

Stata
Kreuter, F. and Valliant, R. (2007). "A survey on survey statistics: What is done and can be done in Stata." *The Stata Journal*, 7(1), 1 – 21. Retrieved August 16, 2012 from http://www.stata-journal.com/article.html?article=st0118

The remainder of this section will provide examples using SAS. It is recommended that if the reader is using a software package other than SAS, he/she should refer to their own software's documentation or customer support for details on the use of replicate weights in variance estimation. However, the following examples may facilitate an understanding of how to create the replicate variance estimates in their own software.

NEW TO SAS 9.2

For most estimates described in this document*, SAS 9.2 is able to calculate variance estimates with the given replicate weight file using Fay's BRR method. In addition, t statistics and confidence intervals using the BRR estimates can be invoked. Additional examples are provided in the unabridged verison of this document.

The following syntax shows how one would use the surveymeans procedure and the replicate weight file to generate a standard error for a population total estimate. Input-related items are highlighted in yellow.

```
proc surveymeans data=dataset sum std clsum cvsum varmethod=brr(fay);
 var variable;
 weight wgt90geo;
 repweights repwgt1-repwgt160;
run;
```

(NOTE: The standard error for the total in proc surveymeans is listed in the SAS output as "Std Dev."

* Proc surveymeans does not calculate BRR variance estimates for medians. Support for the use of BRR to estimate the sampling variance of a median can be found in [3].

IF SAS 9.2 IS NOT AVAILABLE

If SAS 9.2 is not available, standard errors can be calculated using data steps and macros. The following syntax can be used to calculate the standard error for a population total.

```
libname sas 'C:\file_pathway';
*************************************************************************;
* The FIRST STEP is to flag the data records desired after creating  *;
* the SAS data sets. This example flags owner occupied housing units *;
*************************************************************************;
data data1;
 merge sas.AHS20XX sas.REPWGT_20XX;  /*ensure each dataset is sorted by control*/
 by control;
ownocc = (STATUS = 1 and TENURE = 1);  /*another way to say if STATUS = 1 and*/
run;                                   /*TENURE=1 then ownocc=1; else ownocc = 0;*/
*************************************************************************;
* The SECOND STEP of code sums the full sample and the 160 replicate*;
* weights and writes them out to a file.                            *;
*************************************************************************;
proc means data=data1 sum noprint;
 where ownocc=1;              /*For owner-occupants, calculate the sums*/
 var wgt90geo repwgt1-repwgt160;     /*of the full sample and the replicates*/
 output out=data2 sum=est rw1-rw160;
run;

*************************************************************************;
* The THIRD STEP of code uses the estimates of the full sample and the *;
* 160 replicates to compute the estimated replicate variance(s) using  *;
* the formula(s) for 160 replicates. SAS 9.2 also presents an          *;
* alternative to the formula.                                          *;
*************************************************************************;
data data3 (keep=est var se lcl ucl cv);
 set data2 end=eof;
 array repwts{160} rw1-rw160;          /*Fill array with the replicate sums*/
 array sdiffsq{160} sdiffsq1-sdiffsq160;  /*Fill array with the squared diffs*/
     do j = 1 to 160;
       sdiffsq{j} = (repwts{j} - est)**2;
     end;
            totdiff = sum(of sdiffsq1-sdiffsq160);    /*Sum the squared diffs*/
                    var = (4/160) * totdiff;
       se = (var)**.5;
            lcl = est - tinv(0.975,160)*se;
            ucl = est + tinv(0.975,160)*se;
       cv = se/est;
       output;
   run;

proc print data = data3;
  var est var se lcl ucl cv;
run;
```

A general structure exists for the replication of most estimates once they are obtained. **Section 1** shows how to obtain the full-sample and replicate estimates in SAS, given a data set and replicate weights. **Section 2** merges the replicate estimates with the full-sample estimate and calculates the BRR variance estimate. Standard errors, confidence intervals, coefficients of variation, and statistical tests can also be calculated.

The following example calculates a standard error for a mean. By using different statistical procedures, the syntax in Section 1 can be adapted for other types of statistics. Examples are provided in the unabridged version.

Section 1 – General Structure to Obtain Replicate Estimates

```
/*If using this code, include these %let statements. The code references them where applicable*/

%let var = variable;         /*For 1-variable statistics            */
%let dataset = dataset;    /*Name of the data set used in analysis*/

proc means data=&dataset mean noprint;
var &var;
weight wgt90geo;
output out=meansfull (drop=_type_ _freq_) mean=full;
run;

%macro repmeans;
  %do i=1 %to 160;
    proc means data=&dataset mean noprint;
            weight repwgt&i.;
            var &var;
            output out=meanrep&i (drop=_type_ _freq_) mean=repmean&i;
    run;
  %end;
%mend repmeans;
run;

%repmeans;
run;
```

Section 2 – General Structure to Calculate Variances

```
*************************************************;
* Merge the replicates with the full sample estimate *;
*************************************************;

%macro meanbrr;
        data meanreps; /*Merge will create an array of 1 row with 160 variables*/
                merge
                        %do i=1 %to 160;
                                meanrep&i
                        %end;
                meansfull
                ;
run;
%mend meanbrr;
run;

%meanbrr;
run;

*****************************;
* Calculate the estimates   *;
*****************************;

data se_mean (keep=full var se); /* Modify "keep=" to keep the needed calcs.*/
    set meanreps;       /*Select needed calcs given in the data set.*/
    array repwts{160} repmean1-repmean160;
         array sdiffsq{160} sdiffsq1-sdiffsq160;
    do j = 1 to 160;
      sdiffsq{j} = (repwts{j} - full)**2;
    end;
         totdiff = sum(of sdiffsq1-sdiffsq160);

         var = (4/160) * totdiff;
      se = (var)**.5;
        cv = se/full;

/*Confidence Intervals and hypothesis testing can be manually added as needed*/
*          lcl = full - tinv(0.975,160)*se;
*          ucl = full + tinv(0.975,160)*se;
*          t = full/se;
*          p_val =2*(1- cdf('T',abs(t),160));

      output;
   run;

proc print data=se_mean;
format full 10.5;
run;
```

References

[1] U.S. Census Bureau, American Housing Survey for the United States: 2009 http://www.census.gov/hhes/www/housing/ahs/nationaldata.html

[2] Judkins, D. (1990) "Fay's Method for Variance Estimation," Journal of Official Statistics, Vol. 6, No. 3, 1990, pp.223-239.

[3] Thompson, K.J. and Sigman, R.S. (2000), "Estimation and Replicate Variance Estimation of Median Sales Prices of Sold Houses," Survey Methodology, Vol. 26, No. 2, pp. 153-162.

[4] Wolter, Kirk (1985), Introduction to Variance Estimation, New York: Springer-Verlag New York Inc.

[5] Fay, Robert, and Train, George (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," Proceedings of the Section on Government Statistics, American Statistical Association, Alexandria, VA, pp. 154-159.

[6] Plackett, R.L. and Burman, J.P. (1946), "The Design of Optimal Multifactorial Experiments," Biometrika, 33, pp. 305-325.

[7] SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc.