Paper 970-2017

Not So Simple: Intervals You Can Have Confidence in with Real Survey Data

David A. Vandenbroucke, U.S. Department of Housing and Urban Development[1]

ABSTRACT

Confidence intervals are critical to understanding your survey data. If your intervals are too narrow, you might inadvertently judge a result to be statistically significant when it is not. While many familiar SAS® procedures, such as PROC MEANS and PROC REG, provide statistical tests, they rely on the assumption that the data come from a simple random sample. However, almost no real-world survey uses such sampling. Learn how to use the SURVEYMEANS procedure and its "SURVEY" cousins to estimate confidence intervals and perform significance tests that account for the structure of the underlying survey, including the replicate weights now supplied by some statistical agencies. Learn how to extract the results you need from the flood of output that these procedures deliver

INTRODUCTION

This paper describes the use of replicate weights and other methods to more accurately estimate standard errors in survey data. It is aimed at people who want to analyze survey data but are not professional statisticians. It emphasizes how to use the SAS tools, and not on the statistical theory behind them. It begins by reviewing the most basic concepts of statistical inference. Then it describes the differences between simple random samples and the sampling designs that most real-world surveys use. It provides a quick graphical illustration of why this matters. Then it describes how to use SAS survey procedures, concentrating on PROC SURVEYMEANS. The main discussion is about making use of replicate weights, but it also discusses some things you can do if you don't have such weights. It concludes with a few pointers about how to put PROC SURVEYMEANS output into more useful form.

BASIC STATISTICAL INFERENCE

Let's start by reviewing the framework of significance testing in classical, relative frequency, statistics. You have some kind of random variable with a hypothetical central value and a probability distribution around it. This could be a population mean, a regression coefficient, a proportion, a sum, a frequency, any other quantity of interest. The hypothetical value could be the conventional null hypothesis that the parameter is equal to zero; a policy threshold that may or may not be crossed; or the central value of a known population. You also have a point estimate from survey data, and you want to know whether this point estimate is sufficiently different from the hypothetical central value that you can be confident the difference could not be caused by random variation.

The way you decide, or infer, this is you bracket the center of the distribution so that the area in the brackets has a good dose of the total probability—typically 90 or 95 percent. The bracket is called a confidence interval. In normal and similar distributions, a 95 percent bracket is about two standard deviations above and below the mean. If your sample estimate is inside this bracket, you conclude that it's not significantly different. If it's out on one of these edges, you conclude that it is.

Figure 1 is an example using the data from the 2013 American Housing Survey.[2] This shows a normal distribution of the mean of monthly housing cost. The curve is centered on the sample mean, which is $1136 per month. The curve is based on the standard error as computed by PROC MEANS, which works out to about 3.8. This yields a 95 percent confidence interval between $1130 and $1145 per month. Thus, if you observe the mean housing cost for a subgroup, you could infer that it is not statistically

[1] All views expressed in this paper are those of the author and should not be considered to be official positions of the United States government.
[2] http://www.census.gov/programs-surveys/ahs.html

different from the population mean if it was in the range, $1136 plus or minus $7. Clearly, there are more sophisticated ways to model housing cost, but the principle of inference is the same.

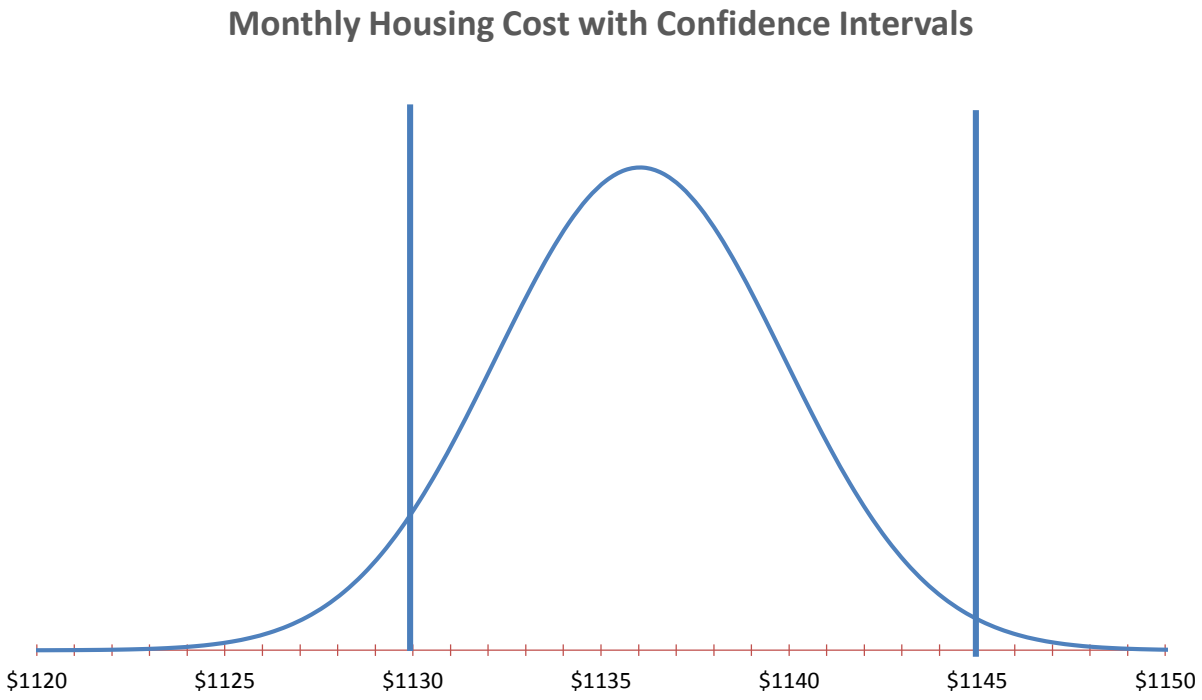

Figure 1. Basic Statistical Inference

Note that making the correct inference depends on accurately measuring the width of the confidence interval. Other ways to say the same thing are to accurately estimate the standard error or standard deviation of your data. The trouble is, if you rely on PROC MEANS to do this, you're in danger of getting it wrong. PROC MEANS is designed to work with simple random samples. In a simple random sample, every member of the population has an equal probability of being selected. Each case in the sample dataset has an equal weight. If you draw a one in 2000 sample, then every sample case represents 2000 members of the population—no more, and no less.

STRUCTURED PROBABILITY SAMPLING

Almost all large-scale social and economic surveys use structured probability samples, rather than simple random samples. Examples include the American Community Survey, the Current Population Survey, and the American Housing Survey. There are several reasons why they do not use simple random samples. One is that it is difficult to obtain a complete roster of the target population so that you can randomly select a specified proportion of them. If a survey uses interviewers to collect data, the sample has to be structured to make the best use of its workforce, and this may mean distributing case assignments more evenly than a random draw would give you. Survey sponsors often have an interest in subpopulations, and this may require oversampling certain groups to provide sufficient sample size for small group estimates. At the back end, survey results are often reweighted to account for unit nonresponse or to match control totals taken from external sources, such as the decennial census or population projections.

Clustered sampling, stratified sampling, oversampling, and post-collection adjustments result in datasets whose cases do not have the equal weights. In a structured sample, some cases may represent only small number of population members, while other cases may represent many more than the average sampling rate. In the 2013 American Housing Survey, the median weight is 2,006, and the interquartile range runs from 1,900 to 2,630. However, the minimum weight is 8, and the maximum is 38,474.

SAS makes it easy to produce point estimates from weighted datasets. All you have to do is include a WEIGHT statement in your procedure code where you specify the variable holding the weights. However, the weight statement does not affect standard errors, confidence intervals, or other measures of variability. This is where you can get into trouble. Standard errors calculated using methods appropriate to simple random samples will underestimate the standard errors of complex samples. In other words, if you take the variances or confidence intervals at face value, you will be overconfident about the estimates and less aware of the expected range in your data. You will be tempted to judge differences as statistically significant, when they are not.

Let's take a look at the difference this makes in the example above. In that example, the mean monthly housing cost was $1136, plus or minus $7 with 95 percent confidence. The point estimate is not affected by accounting for the unequal weighting, but the confidence interval is. Using more accurate estimation, the confidence interval is plus or minus $11, or $1125 to $1146. As you can see in Figure 2, the probability distribution is flatter, with longer tails. The regions between the blue vertical lines and the orange ones would have erroneously been considered significantly different from the mean when using PROC MEANS, which is designed for analyzing simple random samples.

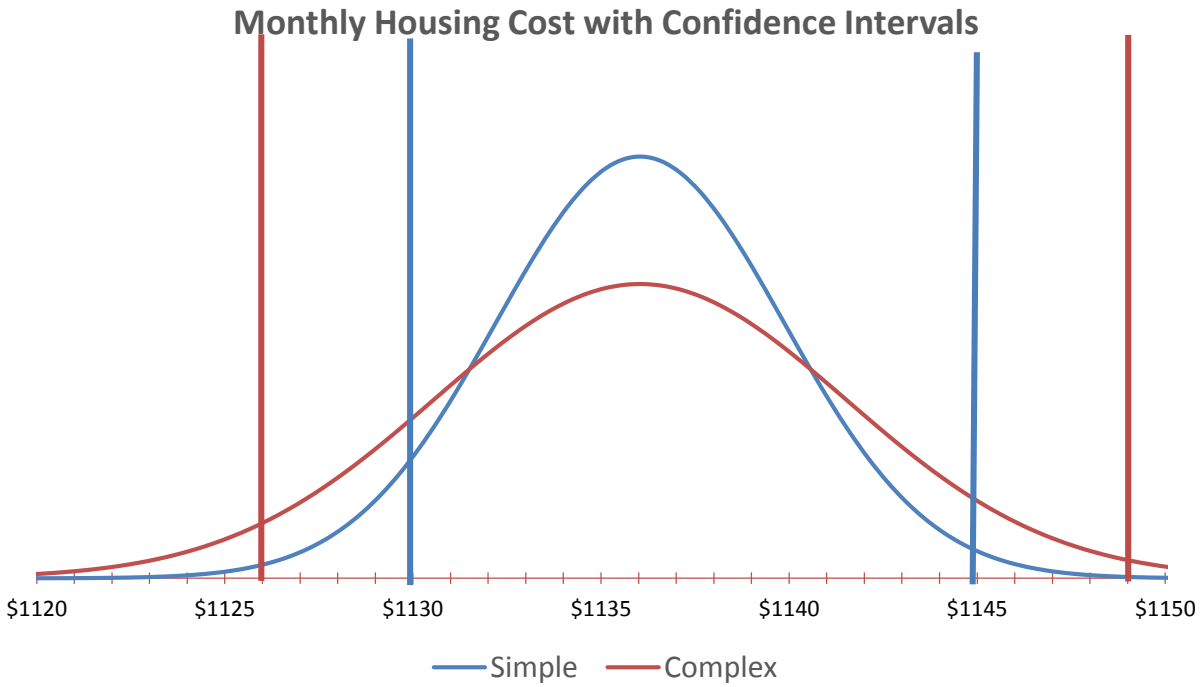

Figure 2 Statistical Inference with a complex sample design

USING SAS SURVEY PROCEDURES

How *do* you analyze data with unequal weights? SASSTAT includes a set of procedures designed to do that. Their names are patterned after the familiar procedures in BASE SAS, except that they begin with "survey:"

- SURVEYMEANS: Descriptive statistics
- SURVEYFREQ: Frequency distributions
- SURVEYREG: Linear regression
- SURVEYLOGISTIC: Logistic regression
- SURVEYPHREG: Proportional hazards modeling
- SURVEYIMPUTE: Imputation of missing values
- SURVEYSELECT: Selection of probability-based samples.

In order to make use of the sample design, you have to use information about it. There are two ways you can do this. The easy way is to use the replicate weights supplied by the organization publishing the survey data. The harder, and often less accurate, way is to identify survey's primary sampling units (PSUs) and the variables that were used to stratify the sample. You can sometimes get this information from survey documentation.

REPLICATE WEIGHTS

Survey datasets include weights that allow you to estimate population values from sample cases. However, the weights associated with the data you have are just one set out of a universe of weights that would have resulted from slightly different samples. If you had more of those samples—a sample of samples, as it were—you would have a better idea of how much variability there was in your data. While this description might give statisticians headaches, that is what replicate weights give you. Replicate weights are re-estimations of the sample weights as if the survey were a little different.

More and more survey organizations are recognizing the importance of providing accurate measures of variability around their estimates. You may have noticed that the US Census Bureau routinely includes a "plus or minus number" in its statistical tables. Some large government surveys that supply replicate weights with their datasets include:

- American Community Survey
- Current Population Survey
- National Crime Victimization Survey
- Consumer Expenditure Survey
- American Housing Survey

What do you see in a dataset that has replicate weights? If you have a dataset with weights, they appear as a variable, usually called something, such as WEIGHT, WGT, or SAMPWT. A dataset with replicate weights contains more variables fields instead of just one: WEIGHT, REPWGT1, REPWGT2, etc., up to some maximum. Thus, replicate weights are just additional variables in your dataset. The naming convention varies, of course. Some surveys call their base weight WEIGHT0, and the replicate weights start with 1 from there. The number of replicate weights will vary from survey to survey. The American Housing Survey public use file supplies160 them, named REPWGT1-REPWGT160.

Where do replicate weights come from? Well, you don't have to worry about the details of that, except that you need to know the general method. SAS survey procedures support replicate weights produced by these methods:

- Jackknife
- Bootstrap
- Balanced Repeated Replication (BRR)

While they are implemented differently, the idea behind all of them is that they extract subsamples from the dataset and calculate weights for each.

Let's see how this works in SAS. First, we run very basic PROC MEANS code as a basis of comparison:

```
/* basic invocation of means */
PROC MEANS DATA=SASGF17.DEMO N MEAN STDERR CLM;
     TITLE Monthly Housing Cost Using PROC MEANS;
     VAR ZSMHC;
     WEIGHT Weight;
RUN;
```

We have one variable ZSMHC, which is monthly housing cost, and we're using a weight variable called WEIGHT. We're asking for the number of observations, the mean, the standard error, and the confidence interval around the mean. The output is shown in Figure 3. There are about 60 thousand cases in the dataset, that the mean cost is $1137.67 per month, and the standard error is 3.8. The confidence interval, which defaults to 95 percent, is plus or minus $7.52 around the point estimate.

Monthly Housing Cost Using PROC MEANS

The MEANS Procedure

Analysis Variable : ZSMHC

N	Mean	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
60097	1137.67	3.8341194	1130.15	1145.18

Figure 3 PROC MEANS output for monthly housing cost.

Now let's do the same thing with PROC SURVEYMEANS, asking for just the default statistics:

```
/* basic invocation of surveymeans */
PROC SURVEYMEANS DATA=SASGF17.DEMO VARMETHOD=BRR(FAY);
     TITLE Monthly Housing Cost Using Replicate Weights;
     VAR ZSMHC;
     WEIGHT Weight;
     REPWEIGHTS RepWgt1-RepWgt160;
RUN;
```

The syntax is very similar. The PROC statement has a VARMETHOD option, which specifies the method used to produce the replicate weights. The AHS uses the balanced repeated replications method, or BRR. It uses a variation on that method called Fay's method, and so that is specified in parentheses. The second difference from PROC MEANS is the REPWEIGHTS statement, which specifies the variables containing the replicate weights, RepWgt1 through RepWgt160.

If we run the code, we get the output shown in Figure 4. The format of the statistics is the same as in Figure 3. Note that the value of the mean is the same. The two procedures will give you the same point estimates. The standard error is shown as about 6.0, higher than the 3.8 computed by PROC MEANS. This is the first difference from the MEANS output—the standard error is higher. Looking at the confidence interval, we see that it is wider: plus or minus $11.92. This illustrates the point made above. After accounting for the sample design there is a wider area of uncertainly. If rely on PROC MEANS, you will be overly confident in the precision of your estimates.

Notice that there was very little extra effort to run PROC SURVEYMEANS instead of MEANS. All we added was an option on the PROC statement and statement to list the replicate weight variables.

Statistics

Variable	N	Mean	Std Error of Mean	95% CL for Mean	
ZSMHC	60097	1137.668117	6.034896	1125.74979	1149.58644

Figure 4 PROC SURVEYMEANS output for monthly housing cost.

ANALYZING SUBGROUPS

Most of the time, you will be interested in subgroups of the population under study, to see how they differ. In surveys of people, these often involve age, sex, race, and so on. When using PROC MEANS you have a number of choices, such as using a CLASS statement or a BY statement. If you are using replicate weights, you need to analyze the entire sample dataset in order to get accurate confidence intervals. This means that you can't use a BY statement, because that divides the data into subpopulations before the statistics are calculated. Instead, SAS provides the DOMAIN statement:

```
/* Using the Domain Statement */
PROC SURVEYMEANS
     DATA=DEMO VARMETHOD=BRR(FAY);
     VAR ZInc2 ZSMHC;
     WEIGHT Weight;
     REPWEIGHTS repwgt1-repwgt160;
     FORMAT Tenure $tenure.;
     DOMAIN Tenure;
RUN;
```

In this code add a second analysis variable, Zinc2, which is household income. It asks for analysis of the subgroups defined by tenure. The output is shown in Figures 5. We see the statistics of both household income and monthly housing costs for owner-occupied housing, cash rent housing, and housing that is occupied without payment of rent.

Statistics for TENURE Domains

TENURE	Variable	N	Mean	Std Error of Mean	95% CL for Mean	
Owned	ZINC2	35852	81701	496.098550	80721.4751	82680.9667
	ZSMHC	35852	1243.358974	7.781910	1227.9905	1258.7275
Cash Rent	ZINC2	23358	43079	409.966997	42269.4400	43888.7289
	ZSMHC	23358	974.308846	5.987174	962.4848	986.1329
No-Cash Rent	ZINC2	887	34545	1356.076580	31867.0499	37223.2854
	ZSMHC	887	170.338987	6.167221	158.1593	182.5186

Figure 5 Domain analysis by tenure.

SAS automatically produces a box-and-whiskers plot with domain analysis, as you can see in Figure 6. This is the plot for monthly housing cost. The full-sample distribution is dominated by the cost for owner-occupied units, because they are such a large proportion of the housing units in the United States. The mean cost for rental units is considerably less than for owner-occupied units, and units housing occupied without payment of rent is much lower still.

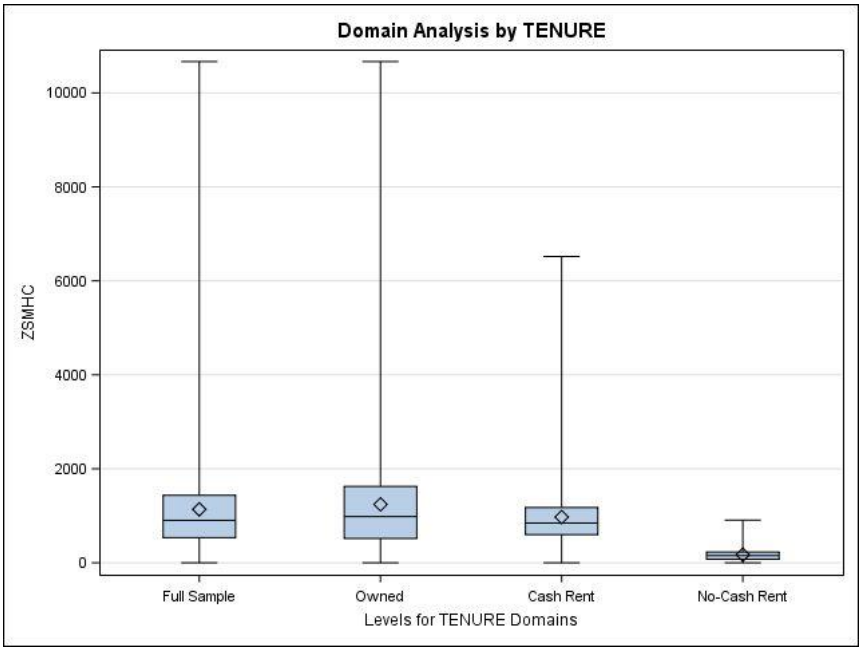

Figure 6 Domain plot by tenure.

Subgroups can be crossed in the usual way, by using an asterisk operator:

```
/* Using the Domain Statement with crossed domains */
PROC SURVEYMEANS
     DATA=SASGF17.DEMO VARMETHOD=BRR(FAY) PLOTS=NONE;
     TITLE Using crossed domains;
     VAR ZSMHC;
     WEIGHT Weight;
     REPWEIGHTS RepWgt1-RepWgt160;
     FORMAT Tenure $tenure. Metro3 $metro.;
     DOMAIN Tenure * Metro3;
RUN;
```

The output is shown in Figure 7. You will note that the PROC statement includes the PLOTS = NONE option. If you don't want the default plots, this will reduce the processing time. Processing time is an issue for the SURVEYxxx procedures, especially when you ask for large numbers of variables and subgroups. The code above took about five minutes to run with the default plots but only 25 seconds without them.

Using crossed domains

The SURVEYMEANS Procedure

Statistics for TENURE*METRO3 Domains

TENURE	METRO3	Variable	N	Mean	Std Error of Mean	95% CL for Mean	
Owned	Central city of MSA	ZSMHC	8173	1325.975530	15.124122	1296.10688	1355.84418
	Suburb - urban	ZSMHC	15170	1502.212486	13.689628	1475.17682	1529.24815
	Suburb - rural	ZSMHC	5314	1211.696984	23.986377	1164.32625	1259.06772
	Outside MSA, urban	ZSMHC	2197	874.270033	21.736874	831.34185	917.19822
	Outside MSA, rural	ZSMHC	4998	809.229736	16.013454	777.60474	840.85473
Cash Rent	Central city of MSA	ZSMHC	11232	999.985809	8.657728	982.88765	1017.08397
	Suburb - urban	ZSMHC	8068	1079.150487	9.508698	1060.37174	1097.92923
	Suburb - rural	ZSMHC	1080	961.408623	19.975921	921.95815	1000.85910
	Outside MSA, urban	ZSMHC	1787	712.340581	14.029558	684.63359	740.04758
	Outside MSA, rural	ZSMHC	1191	693.089096	16.256031	660.98503	725.19316
No-Cash Rent	Central city of MSA	ZSMHC	274	170.145934	13.786223	142.91950	197.37237
	Suburb - urban	ZSMHC	215	199.749650	12.823181	174.42513	225.07417
	Suburb - rural	ZSMHC	111	154.603963	12.333497	130.24652	178.96140
	Outside MSA, urban	ZSMHC	81	191.885508	15.968098	160.35009	223.42093
	Outside MSA, rural	ZSMHC	206	150.110765	9.267962	131.80745	168.41408

Figure 7 Using crossed domains.

WHAT IF YOU DON'T HAVE REPLICATE WEIGHTS?

If you don't have replicate weights, it gets a little harder. You still have to know something about how the sample was drawn, such as how it was clustered by PSUs or stratified by external variables. Survey organizations, such as the U.S. Census Bureau, are usually reluctant to identify their sample cases by PSU, as that can compromise their respondents' confidentiality. Of course, if your data came from a survey that you developed yourself, you will know this information. Most surveys do document their sample designs; this will often tell you what variables were used to stratify the sample. We will look at the two sources of information separately, but you can use both of them if you are lucky enough to have them.

Using Stratification Variables: The STRATA statement

Surveys stratify their samples by using data about their population of interest that is available before data collection. Such data is usually limited to geographic information and demographic aggregates. The American Housing Survey does not stratify within PSUs. Below is an example of how one would make use of stratification if the data were stratified by urban type code, tenure, and mobile home status:

```
/* using stratification variables */
PROC SURVEYMEANS DATA=SASGF17.DEMO PLOTS=NONE;
     TITLE Using stratification variables;
     VAR ZSMHC;
     STRATA Urbcode Tenure Mobile /LIST;
     WEIGHT Weight;
RUN;
```

Note that there is no VARMETHOD option on the PROC statement. The stratification variables are listed on the STRATA statement. SAS will divide the data into strata corresponding to all possible combinations of the variables' values. The /LIST option directs SAS to output a listing of all the strata and the frequency of valid cases in each, for each analysis variable. The stratum information from this example is shown in Figure 8.

Stratum Information

Stratum Index	URBCODE	TENURE	MOBILE	N Obs	Variable	N
1	Central City	Owned	Mobile Home	138	ZSMHC	138
2	Central City	Owned	Not Mobile Home	8035	ZSMHC	8035
3	Central City	Cash Rent	Mobile Home	19	ZSMHC	19
4	Central City	Cash Rent	Not Mobile Home	11213	ZSMHC	11213
5	Central City	No-Cash Rent	Mobile Home	2	ZSMHC	2
6	Central City	No-Cash Rent	Not Mobile Home	272	ZSMHC	272
7	Suburb	Owned	Mobile Home	781	ZSMHC	781
8	Suburb	Owned	Not Mobile Home	19703	ZSMHC	19703
9	Suburb	Cash Rent	Mobile Home	197	ZSMHC	197
10	Suburb	Cash Rent	Not Mobile Home	8951	ZSMHC	8951
11	Suburb	No-Cash Rent	Mobile Home	27	ZSMHC	27
12	Suburb	No-Cash Rent	Not Mobile Home	299	ZSMHC	299
13	Nonmetro	Owned	Mobile Home	725	ZSMHC	725
14	Nonmetro	Owned	Not Mobile Home	4273	ZSMHC	4273
15	Nonmetro	Cash Rent	Mobile Home	160	ZSMHC	160
16	Nonmetro	Cash Rent	Not Mobile Home	1031	ZSMHC	1031
17	Nonmetro	No-Cash Rent	Mobile Home	37	ZSMHC	37
18	Nonmetro	No-Cash Rent	Not Mobile Home	169	ZSMHC	169

Figure 8 Stratification example.

Using PSUs: The CLUSTER statement

If you do know the PSUs that your data came from, perhaps because you administered the survey yourself, then you can use the CLUSTER statement to specify a variable that identifies those to the procedure. If you use the OUTWEIGHTS= option on the PROC statement, you can output replicate weights to a dataset for later use or sharing. In the example below, the variable PSUCode identifies the PSUs:

```
/* using cluster variables */
PROC SURVEYMEANS DATA=SASGF17.DEMO PLOTS=NONE
     VARMETHOD=BRR(FAY) OUTWEIGHTS=SASGF17.RepWeights
     ;    /* end of proc */
     TITLE Using stratification variables;
     VAR ZSMHC;
     CLUSTER PSUCode;
     WEIGHT Weight;
RUN;
```

ODS OUTPUT AND IMPROVED PRESENTATION

The output from PROC SURVEYMEANS can be sent to any ODS destination, just like any other SAS procedure. Even so, you may have noticed that the output table is not very attractive and can be hard to read if you have many analysis variables and domains. Because SURVEYMEANS must act on the entire survey dataset, you may get output for domains that aren't relevant to your work. One way to improve the presentation is to send the output to a SAS dataset and then use a reporting procedure to format the results.

You send output to a dataset using the ODS OUTPUT statement. The procedure is capable of creating a variety of output tables. The one we're interested in is the DOMAIN table:

```
/* Sending domain results to output dataset */
PROC SURVEYMEANS
     DATA=SASGF17.DEMO VARMETHOD=BRR(FAY) PLOTS=NONE;
     TITLE Using Replicate Weights and ODS output to a dataset;
     VAR ZInc2 ZSMHC;
     WEIGHT Weight;
     REPWEIGHTS RepWgt1-RepWgt160;
     DOMAIN Tenure * Metro3;
     ODS OUTPUT DOMAIN=SASGF17.DEMODomain;
RUN;
```

VIEWTABLE: Sasgf17.Demodomain (Statistics for TENURE*METRO3 Domains)

	DomainLabel	TENURE	METRO3	VarName	N	Mean	StdErr	LowerCLMean	UpperCLMean
1	TENURE*METRO3	Owned	Central city of MSA	ZINC2	8173	82640	1111.177538	80445.3427	84834.2752
2	TENURE*METRO3	Owned	Central city of MSA	ZSMHC	8173	1325.975530	15.124122	1296.1069	1355.8442
3	TENURE*METRO3	Owned	Suburb - urban	ZINC2	15170	94385	950.355439	92508.2297	96261.9466
4	TENURE*METRO3	Owned	Suburb - urban	ZSMHC	15170	1502.212486	13.689628	1475.1768	1529.2482
5	TENURE*METRO3	Owned	Suburb - rural	ZINC2	5314	84146	1870.864862	80450.7819	87840.3298
6	TENURE*METRO3	Owned	Suburb - rural	ZSMHC	5314	1211.696984	23.986377	1164.3263	1259.0677
7	TENURE*METRO3	Owned	Outside MSA, urban	ZINC2	2197	62306	1654.073222	59039.3388	65572.6024
8	TENURE*METRO3	Owned	Outside MSA, urban	ZSMHC	2197	874.270033	21.736874	831.3418	917.1982
9	TENURE*METRO3	Owned	Outside MSA, rural	ZINC2	4998	60968	1075.820164	58843.5454	63092.8232
10	TENURE*METRO3	Owned	Outside MSA, rural	ZSMHC	4998	809.229736	16.013454	777.6047	840.8547
11	TENURE*METRO3	Cash Rent	Central city of MSA	ZINC2	11232	42794	601.850572	41605.5356	43982.7269
12	TENURE*METRO3	Cash Rent	Central city of MSA	ZSMHC	11232	999.985809	8.657728	982.8876	1017.0840
13	TENURE*METRO3	Cash Rent	Suburb - urban	ZINC2	8068	47948	711.199329	46543.6718	49352.7691
14	TENURE*METRO3	Cash Rent	Suburb - urban	ZSMHC	8068	1079.150487	9.508698	1060.3717	1097.9292
15	TENURE*METRO3	Cash Rent	Suburb - rural	ZINC2	1080	46726	1605.105069	43555.9786	49895.8276
16	TENURE*METRO3	Cash Rent	Suburb - rural	ZSMHC	1080	961.408623	19.975921	921.9581	1000.8591
17	TENURE*METRO3	Cash Rent	Outside MSA, urban	ZINC2	1787	31751	1531.918694	28725.4168	34776.1940
18	TENURE*METRO3	Cash Rent	Outside MSA, urban	ZSMHC	1787	712.340581	14.029558	684.6336	740.0476
19	TENURE*METRO3	Cash Rent	Outside MSA, rural	ZINC2	1191	33917	1264.521688	31419.7936	36414.4053
20	TENURE*METRO3	Cash Rent	Outside MSA, rural	ZSMHC	1191	693.089096	16.256031	660.9850	725.1932
21	TENURE*METRO3	No-Cash Rent	Central city of MSA	ZINC2	274	29646	2333.290846	25038.0354	34254.0749
22	TENURE*METRO3	No-Cash Rent	Central city of MSA	ZSMHC	274	170.145934	13.786223	142.9195	197.3724
23	TENURE*METRO3	No-Cash Rent	Suburb - urban	ZINC2	215	38634	2611.232639	33476.6700	43790.5248
24	TENURE*METRO3	No-Cash Rent	Suburb - urban	ZSMHC	215	199.749650	12.823181	174.4251	225.0742
25	TENURE*METRO3	No-Cash Rent	Suburb - rural	ZINC2	111	40249	4550.995206	31261.5728	49237.1079
26	TENURE*METRO3	No-Cash Rent	Suburb - rural	ZSMHC	111	154.603963	12.333497	130.2465	178.9614
27	TENURE*METRO3	No-Cash Rent	Outside MSA, urban	ZINC2	81	25315	2523.398028	20331.3173	30298.2427
28	TENURE*METRO3	No-Cash Rent	Outside MSA, urban	ZSMHC	81	191.885508	15.968098	160.3501	223.4209
29	TENURE*METRO3	No-Cash Rent	Outside MSA, rural	ZINC2	206	35770	2835.455495	30170.4776	41369.9686
30	TENURE*METRO3	No-Cash Rent	Outside MSA, rural	ZSMHC	206	150.110765	9.267962	131.8075	168.4141

Figure 9 Domain output dataset.

In this example, we've added a second analysis variable, ZInc2, which is household income. We are generating statistics for housing tenure crossed with metropolitan status. A screenshot of the dataset is shown in Figure 9. The analysis variables are interleaved in much the same way that they would be in printed output. However, now that we have the results in a dataset, we can use PROC TABULATE to format them in a more reader-friendly way. First we define a format to make the variable listing more readable:

```
PROC FORMAT;
     VALUE $Varname
          'ZSMHC' = 'Monthly Housing Cost'
          'ZINC2' = 'Household Income'
     ; /* end of value statement */
RUN;
```

Then we set up an ODS destination to send the output to an Excel spreadsheet:

```
ODS EXCEL FILE='J:\path\ DemoDomain.xlsx'
      OPTIONS(EMBEDDED_TITLES='YES');
```

The PROC TABULATE step begins by renaming some of the variables in the DemoDomain dataset (MEAN, STDERR, and N), because TABULATE would see them as statistics keywords. The crossed domain variables, TENURE and METRO3, are used to order the rows. VARNAME contains the names of the analysis variables. It is used to order the columns, with the format applied. Each set of statistics is crossed with its corresponding analysis variable. The procedure requests a SUM statistic. However, given that there is only one row in the dataset for each combination, this is essentially a detail report:

```
PROC TABULATE
     DATA = SASGF17.DemoDomain
          (RENAME=(Mean=MeanVar StdErr=StdErrVar N=Cases))
          ORDER=DATA
     ; /* end of proc statement */
     TITLE Household Income and Monthly Housing Cost by Metro Status
     and Tenure, 2013;
     CLASS Tenure Metro3 VarName;
     VAR  Cases MeanVar StdErrVar LowerCLMean UpperCLMean;
     FORMAT VarName $varname.;
     TABLE
          (Metro3='Metro Status' * Tenure='Tenure'),    /* rows */
          VarName='Analysis Variables'*                 /* columns */
               (
                         Cases          ='Sample Cases'
                         MeanVar        ='Mean'
                         StdErrVar      ='Standard Error'
                         LowerCLMean    ='Lower Confidence Limit'
                         UpperCLMean    ='Upper Confidence Limit'
                )*SUM=' '*F=COMMA6.     /* "sums" over one record each*/
               / PRINTMISS MISSTEXT= ' '
     ;    /* end of table */
RUN;
ODS EXCEL CLOSE;
```

The spreadsheet is shown in Figure 10.

CONCLUSION

While this paper has focused on mean values and PROC SURVEYMEANS, the syntax for other statistics is very similar. SURVEYMEANS can also produce estimates of and ratios among analysis variables by domain. The other SURVEYxxx procedures for regression, frequencies, logistic regression, and survival analysis use syntax to their corresponding procedures in BASE SAS. The survey-related statements are the same as have been described in this paper. Thus, you can easily adapt your current work to produce more accurate standard errors, confidence intervals, and other measures of variability.

Household Income and Monthly Housing Cost by Metro Status and Tenure, 2013

		Analysis Variables									
		Household Income					Monthly Housing Cost				
Metro Status	Tenure	Sample Cases	Mean	Standard Error	Lower Confidence Limit	Upper Confidence Limit	Sample Cases	Mean	Standard Error	Lower Confidence Limit	Upper Confidence Limit
Central city of MSA	Owned	8,173	82,640	1,111	80,445	84,834	8,173	1,326	15	1,296	1,356
	Cash Rent	11,232	42,794	602	41,606	43,983	11,232	1,000	9	983	1,017
	No-Cash Rent	274	29,646	2,333	25,038	34,254	274	170	14	143	197
Suburb - urban	Owned	15,170	94,385	950	92,508	96,262	15,170	1,502	14	1,475	1,529
	Cash Rent	8,068	47,948	711	46,544	49,353	8,068	1,079	10	1,060	1,098
	No-Cash Rent	215	38,634	2,611	33,477	43,791	215	200	13	174	225
Suburb - rural	Owned	5,314	84,146	1,871	80,451	87,840	5,314	1,212	24	1,164	1,259
	Cash Rent	1,080	46,726	1,605	43,556	49,896	1,080	961	20	922	1,001
	No-Cash Rent	111	40,249	4,551	31,262	49,237	111	155	12	130	179
Outside MSA, urban	Owned	2,197	62,306	1,654	59,039	65,573	2,197	874	22	831	917
	Cash Rent	1,787	31,751	1,532	28,725	34,776	1,787	712	14	685	740
	No-Cash Rent	81	25,315	2,523	20,331	30,298	81	192	16	160	223
Outside MSA, rural	Owned	4,998	60,968	1,076	58,844	63,093	4,998	809	16	778	841
	Cash Rent	1,191	33,917	1,265	31,420	36,414	1,191	693	16	661	725
	No-Cash Rent	206	35,770	2,835	30,170	41,370	206	150	9	132	168

Figure 10 Domain Statistics Displayed with PROC TABULATE.

RECOMMENDED READING

- *American Housing Survey* https://www.huduser.gov/portal/datasets/ahs.html
- Lewis, Taylor. . Replication Techniques for Variance Approximation. Joint Program in Survey Methodology (JPSM), Paper 2601-2015. University of Maryland. https://support.sas.com/resources/papers/proceedings15/2601-2015.pdf
- *SAS/STAT® 14.2 User's Guide The SURVEYMEANS Procedure* http://support.sas.com/documentation/onlinedoc/stat/142/surveymeans.pdf

CONTACT INFORMATION

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

Dav Vandenbroucke
U.S. Department of Housing and Urban Development
202-402-5890
david.a.vandenbroucke@hud.gov