Paper AA-09-2013

Analyzing Continuous Variables from Complex Survey Data Using PROC SURVEYMEANS

Taylor Lewis, University of Maryland, College Park, MD

ABSTRACT

This paper explores features available in PROC SURVEYMEANS to analyze continuous variables in a complex survey data set, where "complex" denotes a data set characterized by one or more of the following features: unequal weights, stratification, clustering, and finite population corrections. Using a real-world complex survey data set, this paper demonstrates the necessary syntax to have PROC SURVEYMEANS properly estimate totals, means, ratios, and quantiles, as well as their corresponding design-based measures of variability.

INTRODUCTION

This paper explores the capabilities of PROC SURVEYMEANS for analyzing variables that can effectively be treated as continuous. Discussion and syntax examples center around publically-available data from the 2003 Commercial Buildings Energy Consumption Survey (CBECS)—background given below—which teems with continuous variables. Examples include square footage of the sampled building, its kilowatt-hours of electricity consumed, and any associated expenditures. A limited amount of consideration is given to dichotomous or nominal variables, which can technically be analyzed in PROC SURVEYMEANS using the CLASS statement or by creating one or more 0/1 indicator variables (i.e., a unit is assigned a value of 1 if it has the given characteristic, and 0 otherwise).

The paper conforms to a structure whereby a separate section is devoted to each of the four broad classes of statistics available in PROC SURVEYMEANS: totals, means, ratios, and quantiles. Each estimator is first introduced algebraically along with a brief mention of how PROC SURVEYMEANS estimates its variance, followed by one or more elementary syntax examples. Each section concludes with a tabular summarization of the statistical keywords available in the PROC statement. Before launching into these details, however, the paper begins with a brief background section on features of complex survey data as well as a brief background section on the CBECS.

BACKGROUND ON FEATURES OF COMPLEX SURVEY DATA

There are four distinct features of "complex" survey data that can arise:

- Finite population corrections
- Clustering
- Stratification
- Unequal weights

In general, if the data emanate from a sample design that introduced one or more of these features, you should employ a SAS/STAT® analysis procedure prefixed by SURVEY. There are currently five such procedures:

- PROC SURVEYMEANS
- PROC SURVEYFREQ
- PROC SURVEYREG
- PROC SURVEYLOGISTIC
- PROC SURVEYPHREG

All five share a common syntax structure to inform SAS® of these features in the input data set.

In many introductory statistics courses, the implied data collection mechanism is simple random sampling with replacement (SRSWR), possibly from an infinite or hypothetical population. Under that paradigm, data are assumed independently and identically distributed, or i.i.d. for short. In contrast, survey researchers often select samples without replacement from finite, or enumerable, populations, and simple random sampling is the exception rather than the rule. Alternative sample designs can yield efficiencies in many circumstances, but they are most often pursued out of necessity or to save on data collection costs.

For sake of an example, assume a state board of education is interested in measuring the mathematical aptitude of $N = 1{,}000$ students at a particular high school by way of a standardized test. That is, the finite population of interest is the student body of the given school. Instead of administering the test to all students, suppose a sample of $n = 200$ students is selected

randomly and that an aptitude y_i is measured for each. We know from standard statistical theory that the sample mean is $\hat{\bar{y}} = \frac{\sum_{i=1}^{n} y_i}{n}$ an unbiased estimate of $\bar{y}$, the true population mean, or the average test score for all students in the high school. If the sample were selected with replacement, meaning each student in the population could be sampled (and measured via the test) more than once, the estimated variance of the sample mean would be calculated as $\mathrm{var}(\hat{\bar{y}}) = \frac{1}{n} \frac{\sum_{i=1}^{n} \left(y_i - \hat{\bar{y}}\right)^2}{(n-1)}$. If the sample were selected without replacement, however, the estimated variance formula would more accurately be $\mathrm{var}(\hat{\bar{y}}) = \frac{1}{n} \frac{\sum_{i=1}^{n} \left(y_i - \hat{\bar{y}}\right)^2}{(n-1)} \left(1 - \frac{n}{N}\right)$. In other words, sampling without replacement reduces the variance in proportion to the sampling rate—in this case, 20%.

The term (1 – *n* / *N*) is called the *finite population correction*, or FPC, and enters other estimators' variance formulas, not strictly that of the sample mean. Notice that as the sampling rate *n* / *N* approaches 1, the variance tends to 0, which is an intuitive result. Another way of conceptualizing this is that, as the portion of the population sampled increases, uncertainty in a sample-based estimate decreases. In the most extreme case of a census (when *n* = *N*), the FPC is 0 and there is no variance. The sample-based estimate defaults to the given population quantity.

One straightforward way to incorporate the FPC is to use the TOTAL= option in the PROC statement. SAS determines *n* from the input data set, but relies on the user to specify *N*. Alternatively, you can specify the sampling rate *n* / *N* using the RATE= option. If neither the TOTAL= or RATE= options is present, the SURVEY procedure assumes sampling was conducted with replacement and ignores the FPC.

The second feature is *clustering*, which occurs when the unit sampled is actually a cluster of population units. Returning to our hypothetical example, suppose that each student in the high school starts his or her school day in a homeroom where attendance is taken and other administrative matters handled. For numeric concreteness, assume there are 40 homerooms, each comprised of 25 students. From the standpoint of data collection logistics, it would be much easier to sample homerooms and administer the test therein as opposed to tracking down each sampled student independently. One could still achieve a sample size of 200 by sampling 8 of the 40 homerooms. This is a legitimate sample design, but the clustering should be accounted in the analysis stage by specifying a homeroom identifier variable in the CLUSTER statement of the respective SURVEY procedure.

There is no mandate to sample all units within a cluster. For instance, we could have achieved the same sample size by initially selecting 20 homerooms, then selecting 10 students from each at random. This is an example of a multi-stage sampling design in which the primary sampling units (PSUs) are homerooms and the secondary sampling units (SSUs) are students. It is worth emphasizing, however, that only the PSU identifier should be specified in the CLUSTER statement. When SAS sees two variables in the CLUSTER statement, it assumes the combination of the two defines a PSU, which can result in an unduly low variance estimate. Specifying only the PSU implicitly invokes the *ultimate cluster assumption* (see p. 35 of Kalton, 1983) that is frequently used to simplify variance calculations. A common concern voiced by practitioners is that this does not account for all stages of sampling and, thus, may underestimate variability. More commonly, however, the result is a slight overestimation of variability.

The third feature of complex survey data is *stratification*, which arises when PSUs are allocated into one of a mutually exclusive and exhaustive set of groups, or *strata* (singular: stratum) and an independent sample is selected within each. Whereas clustering typically decreases precision, in all but a few rare circumstances, stratification increases precision. The reason is that the overall variance consists of stratum-specific variance estimates summed over all strata. When strata are constructed homogeneously with respect to the principle outcome variable(s), there can be considerable precision gains relative to simple random sampling.

Returning to our hypothetical example, a prudent stratification variable might be grade level. Suppose the 40 homerooms could be grouped into 4 sets of 10, one for each grade level—ninth through twelfth. Figure 1 illustrates how this might look if 2 homerooms were sampled within each grade. Rows correspond to strata, columns to clusters, and a filled-in cell denotes being selected into the sample. If this particular sample design was employed, however, we would need to inform SAS of the grade level identifier by placing it in the STRATA statement of the given SURVEY procedure.

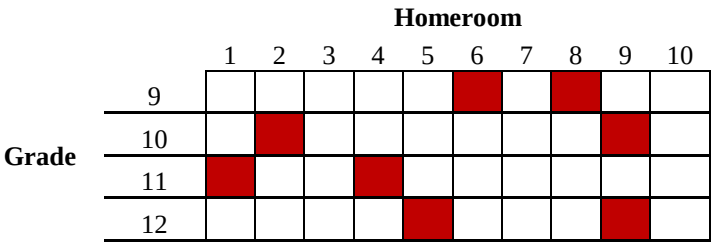

Figure 1. Visual Representation of a Stratified, Cluster Sample for the Hypothetical Mathematics Aptitude Survey

Parenthetically alluded to above was how sampling rates of clusters may vary across strata. In general, when sampling rates vary amongst the ultimately sampled units, one should account for this by assigning a unit-level weight equaling the inverse of that unit's selection probability. Weights are the fourth feature of complex survey data and can be interpreted as the number of population units a sample unit represents. For instance, if a sample unit's selection probability was one-fourth, that unit would be assigned a weight of 4. The unit's survey responses represent itself and three other comparable units in the population. Where applicable, these weights should be stored as a numeric weight variable and specified in the WEIGHT statement of the SURVEY procedure. In the absence of a WEIGHT statement, units are implicitly assigned a weight of 1.

BACKGROUND ON THE CBECS

The Commercial Building Energy Consumption Survey (CBECS) (http://www.eia.gov/consumption/commercial/about.cfm) is sponsored by the Energy Information Administration, a statistical subagency of the U.S. Department of Energy. The sampling unit in this survey is a building. Eligible buildings include those at least 1,000 square feet in size and having more than 50% of its floorspace devoted to activities that are not residential, industrial, or agricultural in nature. Key statistics of interest include characteristics of the building such as square footage, year of construction, types and uses of heating/cooling equipment, and the volume and associated expenditures of energy consumed by the building. One high-profile usage of this survey's data is that it serves as a benchmark for EPA's ENERGY STAR rating system. Specifically, a given building's rating is based on a comparable set of buildings surveyed as part of CBECS with respect to size, location, number of occupants, and other factors. For more information, see http://www.energystar.gov/index.cfm?c=evaluate_performance.pt_neprs_learn. The survey was first administered in 1979 and is generally conducted every four years. In this paper, we will analyze data from the 2003 CBECS available for download at http://www.eia.gov/consumption/commercial/data/2003/index.cfm?view=microdata.

The CBECS sample design begins with a stratified, clustered area probability sample. PSUs are counties or groups of counties, and subsequent stage sampling units are at finer levels of geographic detail. Ultimately, buildings in the finest applicable geographical unit sampled were listed and a random sample of them taken. The area sampling approach is augmented by a few sample frames developed from specialized lists acquired for certain classes of buildings such as those devoted to Federal Government activities, colleges and universities, and hospitals. Data on the sampled building is collected onsite by interviews with the building owners, managers, or tenants, although a supplemental supplier's survey is fielded to capture additional information on energy consumption and expenditures, particularly for onsite respondents who are unable to satisfactorily provide this detailed information.

TOTALS

The first statistic we will consider is the estimated total for a survey outcome variable y, which can be expressed in the most general sense as

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij} \qquad (1)$$

where H is the number of strata, n_h is the number of PSUs selected from the h^{th} stratum, and m_{hi} is the number of units selected from the n_h^{th} PSU. In essence, the estimated total is the weighted sum of all observations in the data set. The expression is given in "the most general sense" because it assumes the sample design includes stratification, clustering, and unequal weights. The expression simplifies in the absence of any of these features.

At first glance, the estimated variance for this statistic is rather complicated. Specifically, the formula provided in the documentation is

$$\text{var}(\hat{Y}) = \sum_{h=1}^{H} \frac{n_h}{(n_h - 1)} \sum_{i=1}^{n_h} \sum_{i=1}^{m_{hi}} \left(w_{hij} y_{hij} - \frac{\sum_{i=1}^{n_h} \sum_{i=1}^{m_{hi}} w_{hij} y_{hij}}{n_h} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \quad (2)$$

where N_h denotes the number of clusters in the h^{th} stratum's population. This equation is also given in the "most general sense," and can simplify when certain complex survey features do not apply. For instance, the rightmost term in parentheses represents a stratum-specific finite population correction with respect to the number of clusters selected in the first stage of sampling. If the design involved multiple stages of clustering and the ultimate cluster assumption was implicitly invoked, this term would disappear.

One of the key variables measured for all sampled buildings in the CBECS is square footage. This is maintained in the variable SQFT8 on the publically-available data set. (The data producers chose to end all variable names in the 2003 CBECS data set with an "8," because the 2003 administration was the 8th in the survey's history.) As was discussed above, the CBECS sample design involves three features of complex survey data: stratification, clustering, and unequal weights. These can be accounted for in any SURVEY procedure by pointing SAS to the variables STRATUM8, PAIR8, and ADJWT8, respectively. The example below illustrates the syntax necessary to estimate the total square footage of all buildings in the population. (The criteria for determining whether a building is eligible for the survey is rather complex and will not be defined here. More details can be found on the survey's website: http://www.eia.gov/consumption/commercial/)

Specifying the keyword SUM in the PROC statement is all that is needed to estimate the total for any variable listed in the VAR statement and its corresponding standard error, the square root of the quantity in (2). This quantity is labeled "Std Dev" in the output, which is somewhat of a misnomer. A listing of additional statistical keywords available in the PROC statement as they relate to totals will be given in Table 1 towards the end of this section.

```
proc surveymeans data=CBECS_2003 sum;
  strata STRATUM8;
  cluster PAIR8;
  var SQFT8;
weight ADJWT8;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Strata	44
Number of Clusters	88
Number of Observations	5215
Sum of Weights	4858749.82

Statistics

Variable	Sum	Std Dev
SQFT8	71657900522	2227163020

PROC SURVEYMEANS output always begins with a brief summary of the complex survey features identified on the input data set. From the "Data Summary" component of the output, we are informed that the 5,215 observations in the data set CBECS_2003 were found to be spread amongst 88 distinct PSUs in 44 distinct strata, and that the weights sum to approximately 4,858,745. To avoid redundancy, this summary will not be shown in subsequent examples appearing in this paper. The "Statistics" component of the output shows that the estimated total square footage is almost 72 billion with a standard error of 2.2 billion.

Summations are also possible for categorical variables. As an example, the variable STUSED8 indicates whether the building uses district steam for heating purposes. A value of 1 indicates yes, whereas 2 means no. The syntax example below demonstrates how we can estimate the total number of buildings for both conditions. Since the STUSED8 variable is numeric but we want it treated as a nominal categorical variable, it is also specified in the CLASS statement (character variables listed in the VAR statement are treated as categorical by default). For each CLASS statement variable, PROC

SURVEYMEANS constructs a sequence of K binary indicator variables, one for each of the k = 1, 2, …, K unique values ("levels" in SAS terminology), defined as y_{hij} = 1 if the observation falls within that category, and y_{hij} = 0 otherwise. From there, formulas (1) and (2) can be employed.

```
proc format;
  value YESNO
    1='Yes'
    2='No';
run;

proc surveymeans data=CBECS_2003 sum;
  strata STRATUM8;
  cluster PAIR8;
  class STUSED8;
  var STUSED8;
weight ADJWT8;
format STUSED8 YESNO.;
run;
```

Class Level Information

Class Variable	Levels	Values
STUSED8	2	Yes No

Statistics

Variable	Level	Sum	Std Dev
STUSED8	Yes	47106	7602.878471
	No	4811644	239130

We can gather from the output that an estimated total of 47,106 buildings use district steam and 4,811,644 did not. Because all buildings are characterized by one of these two conditions, the sum of the two matches the sum of weights for the entire data set, 4,858,745. In fact, this aggregation can be interpreted as an estimate of the total number of buildings $\hat{N}$ in the population (as of 2003). This is actually an estimate of interest, since it is not known elsewhere (i.e., there is no master list or combined set of lists enumerating all eligible buildings in the population). Measures of variability associated with this particular statistic are not output by default. If they are desired, one work-around would be to apply syntax of the form illustrated in the two examples above naming a user-defined numeric variable equaling 1 for all observations in the data set in the VAR statement. An even simpler approach is to omit the WEIGHT statement, put ADJWT8 in the VAR statement, and specify the statistical keyword SUM in the PROC statement. Another alternative would be to use PROC SURVEYFREQ—see Lewis (2013b).

Table 1 summarizes the useful statistical keywords pertaining to estimated totals that can be requested in the PROC SURVEYMEANS statement. Most are self-explanatory, although a few words are warranted about the complex survey degrees of freedom. Under the default variance estimation procedure, the SURVEY procedures utilize a widely-adopted rule of thumb that the degrees of freedom equal the number of distinct PSUs minus the number of strata. (Alternative variance estimation procedures discussed in Mukhopadhyay et al. (2008) abide by a different set of rules.) This is true for any statistic, even linear models. The important thing to realize is that this is often dramatically smaller than the number of observations minus one, what would be assumed for an SRSWR sample design. For instance, the effective degrees of freedom in the examples above are 88 – 44 = 44 not 5,215 – 1 = 5,214. These rules do not affect the point estimates or their estimated variances, but do affect confidence intervals, since their widths are contingent upon a reference t distribution with said degrees of freedom.

Keyword	Output
SUM	Estimated total of each variable listed in the VAR statement or all levels of each categorical variable listed in the CLASS statement
STD	Standard error of the estimated total, output by default when the SUM keyword is used
VARSUM	Variance of the estimated total
CVSUM	Coefficient of variation (STD divided by SUM) of the estimated total
CLSUM	Confidence interval of the estimated total based on the significance level specified in the ALPHA= option (default is ALPHA=.05) and complex survey degrees of freedom (# PSUs – # strata)

Table 1. Summary of PROC SURVEYMEANS Statement Statistical Keywords Related to Estimated Totals

Before proceeding, it is worth mentioning that there is no built-in mechanism for conducting significance tests on estimated totals. In other words, there is no way to input a null hypothesis total and have PROC SURVEYMEANS compute a *t*-statistic and *p*-value. Of course, all of the essential components are output for the user to do so without much hassle. The same is true for other statistics estimated by PROC SURVEYMEANS, with the exception of the null hypothesis that the true population mean or ratio is 0 (see discussion in the documentation regarding the statistical keyword T). Additional methods and considerations for conducting significance tests on sample means are discussed in Lewis (2013a).

MEANS

The second statistic considered in this paper is the sample mean for an outcome variable *y*, which can be expressed as

$$\hat{\bar{y}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}y_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}} \tag{3}$$

Another name for this estimator is the *weighted mean*. It is the weighted sum of values for the given variable divided by the sum of the weights. The weighted mean complicates the variance estimation task because it is actually a nonlinear function (ratio) of two estimated totals, $\hat{\bar{y}} = \frac{\hat{Y}}{\hat{N}}$, and $\text{var}\left(\frac{\hat{Y}}{\hat{N}}\right) \neq \frac{\text{var}(\hat{Y})}{\text{var}(\hat{N})}$. As noted by Heeringa et al. (2010), a closed-form variance formula such as (2) does not exist. This requires one to make an approximation. A popular method is Taylor series linearization (TSL), which is the default approach used by PROC SURVEYMEANS.

While we will not delve too deeply into the computational details of how SAS carries out this TSL process, it turns out that the resulting TSL variance approximation for (3) is

$$\text{var}(\hat{\bar{y}}) \approx \frac{1}{\hat{N}^2}\left[\text{var}(\hat{Y}) + \left(\frac{\hat{Y}}{\hat{N}}\right)^2 \text{var}(\hat{N}) - 2\times\left(\frac{\hat{Y}}{\hat{N}}\right)\times\text{cov}(\hat{Y},\hat{N})\right] \tag{4}$$

Note that the variance terms in (4) are calculated in the manner demonstrated in the Totals section with respect to the complex sample design and $\text{cov}(\hat{Y},\hat{N})$ represents the covariance of $\hat{Y}$ and $\hat{N}$, also calculated with respect to the complex sample design. Ignoring the FPC, this would be calculated as

$$\text{cov}(\hat{Y},\hat{N}) = \sum_{h=1}^{H}\frac{n_h}{(n_h-1)}\sum_{i=1}^{n_h}\sum_{i=1}^{m_{hi}}\left(w_{hij}y_{hij} - \frac{\sum_{i=1}^{n_h}\sum_{i=1}^{m_{hi}} w_{hij}y_{hij}}{n_h}\right)\left(w_{hij} - \frac{\sum_{i=1}^{n_h}\sum_{i=1}^{m_{hi}} w_{hij}}{n_h}\right) \tag{5}$$

Further note that in the absence of weights—when we effectively assume a uniform weight of 1 for all units in the data set—the variance of $\hat{N}$ is 0 as is any covariance involving $\hat{N}$ and equation (4) defaults to an SRSWR sample mean variance

equation (the traditional, non-complex survey formula shown previously). Again, suffice it to say the introduction of unequal weights complicates the process of computing variances. Thankfully, we have SAS to do the legwork for us.

The example below shows how to estimate the average square footage for all buildings in the CBECS population. The syntax mirrors that seen previously, except the keyword MEAN is now specified in the PROC statement. As shown in the output, the estimated mean is 14,748 and the standard error 625.

```
proc surveymeans data=CBECS_2003 mean;
  strata STRATUM8;
  cluster PAIR8;
  var SQFT8;
weight ADJWT8;
run;
```

Statistics

Variable	Mean	Std Error of Mean
SQFT8	14748	625.460801

Means of categorical variables listed in the CLASS statement are output in the form of *proportions*. In a manner similar to what occurs when totals are requested, PROC SURVEYMEANS constructs a sequence of binary indicator variables defined as 1 if the observation falls within the k^{th} level (k = 1,…,K) and 0 otherwise. From there, the formulas of (3) and (4) are used.

One way of thinking about the estimated proportion for the k^{th} category is that it is the sum of the weights for all observations falling within that category divided by the overall sum of weights. It is not uncommon for textbooks to express the variance of proportions somewhat differently from sample means—indeed, the PROC SURVEYMEANS documentation distinguishes the two. For instance, the estimated variance of a sample proportion $\hat{p}$ in an SRSWR sample design is often written as $\text{var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}$. This expression is the result of an algebraic simplification that can be made when summing squared deviations of a variable consisting of either a 0 or 1 from its mean (i.e., the estimated proportion). Calculating the variance this way is far less computationally intensive. But the "standard" equation for a continuous variable still applies when an appropriately constructed 0/1 indicator variable is at hand.

The example below illustrates these concepts by estimating the proportion of buildings in the CBECS population that use natural gas, a characteristic designated by the variable NGUSED8 in the CBECS_2003 data set. This variable is specified in both the VAR and CLASS statements. From the output, it appears about 52.2% of the buildings use natural gas and 47.8% do not. Note that the standard errors of these two point estimates are equivalent. This is a sensible result that occurs whenever K = 2 (i.e., from a variability perspective, it should not matter which category was assigned the 0 and which was assigned the 1).

```
proc format;
  value YESNO
    1='Yes'
    2='No';
run;

proc surveymeans data=CBECS_2003 mean;
  strata STRATUM8;
  cluster PAIR8;
  class NGUSED8;
  var NGUSED8;
weight ADJWT8;
format NGUSED8 YESNO.;
run;
```

Class Level Information

Class Variable	Levels	Values
NGUSED8	2	Yes No

Statistics

Variable	Level	Mean	Std Error of Mean
NGUSED8	Yes	0.522272	0.028083
	No	0.477728	0.028083

Table 2 summarizes the relevant statistical keywords available for means.

Keyword	Output
MEAN	Estimated mean of each variable listed in the VAR statement or estimated proportion of all levels of each categorical variable listed in the CLASS statement
STDERR	Standard error of the estimated mean, output by default when the MEAN keyword is used
VAR	Variance of the estimated mean
CV	Coefficient of variation (MEAN divided by STDERR) of the estimated mean
CLM	Confidence interval of the estimated mean based on the significance level given in the ALPHA= option (default is ALPHA=.05) and complex survey degrees of freedom (# PSUs - # strata)

Table 2. Summary of PROC SURVEYMEANS Statement Statistical Keywords Related to Estimated Means

RATIOS

The third statistic discussed in this paper is the ratio of two totals, $\hat{Y}$ and $\hat{X}$, which can be expressed as

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} y_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} x_{hij}} \qquad (6)$$

Recall it was noted that the weighted mean is a ratio in which the numerator consists of a variate equaling the weight times the given outcome variable and the denominator is simply the weight itself. It is instructive to see how these two variates can be created and passed to PROC SURVEYMEANS to replicate the weighted mean estimation and inference process. Observe how the output from the syntax example matches the output from the mean square footage example above. The two variates are created on the data set RATIO_EXAMPLE and named NUM for numerator and DEN for denominator. The RATIO statistical keyword in the PROC SURVEYMEANS statement requests the point estimate and standard error for the ratio defined in the RATIO statement. The required syntax calls for separating the numerator variable from the denominator variable with a slash. These two variables should also be listed in the VAR statement; without a VAR statement, SURVEYMEANS outputs statistics for all numeric variables in the input data set not already designated for a complex survey

design feature. Note that the WEIGHT statement is unnecessary since the NUM and DEN variables already account for the weighting.

```
data ratio_example;
  set CBECS_2003;
num=(ADJWT8*SQFT8);
den=ADJWT8;
run;

proc surveymeans data=ratio_example ratio;
  strata STRATUM8;
  cluster PAIR8;
  var num den;
  ratio num / den;
run;
```

Ratio Analysis

Numerator	Denominator	Ratio	Std Err
num	den	14748	625.460801

The exchangeability is attributable to the fact that the variance of any kind of ratio mirrors what we have already seen for the weighted mean. Specifically, the TSL variance approximation for (6) is

$$\mathrm{var}(\hat{R}) \approx \frac{1}{\hat{X}^2}\left[\mathrm{var}(\hat{Y}) + \left(\frac{\hat{Y}}{\hat{X}}\right)^2 \mathrm{var}(\hat{X}) - 2\times\left(\frac{\hat{Y}}{\hat{X}}\right)\times\mathrm{cov}(\hat{Y},\hat{X})\right] \qquad (7)$$

An example ratio estimated in the CBECS is *electricity intensity*, defined as the average amount of electricity expended per square foot of building space. To get this figure, the total amount of kilowatt-hours of electricity consumed (ELCSN8) is divided by the total square footage (SQFT8). The example below illustrates how to estimate this ratio. From the output, we observe this statistic to be approximately 14.86 with standard error 0.42.

```
proc surveymeans data=CBECS_2003 ratio;
  strata STRATUM8;
  cluster PAIR8;
  var ELCNS8 SQFT8;
  ratio ELCNS8 / SQFT8;
weight ADJWT8;
run;
```

Ratio Analysis

Numerator	Denominator	Ratio	Std Err
ELCNS8	SQFT8	14.864030	0.420386

Ratio estimation is discussed at length in some of the classic sampling texts (*c.f.*, Ch. 3 of Lohr, 1999; Ch. 6 of Cochran, 1977). For example, one application that can lead to significant efficiencies is when the total for the denominator is known with certainty, possibly from an external source. This value might be denoted X (without a hat). The idea is to estimate the ratio $\hat{R} = \frac{\hat{Y}}{\hat{X}}$ from the sample and multiply it by X to get a better, more precise estimate of Y. That is, instead of the estimator shown in (6), one would use

$$\hat{Y}_{ratio} = X\hat{R} = X\left(\frac{\hat{Y}}{\hat{X}}\right) \quad (8)$$

with an estimated variance of

$$\text{var}(\hat{Y}_{ratio}) = X^2 \text{var}\left(\hat{R}\right) \quad (9)$$

where $\text{var}\left(\hat{R}\right)$ would be calculated as shown in (7). There is no way to have this particular ratio estimator output directly, but an indirect way would be to create a new numerator variable equaling the original numerator times *X*, and then specify it variable in the RATIO statement alongside the original denominator. From the output of a PROC SURVEYMEANS run comparable to the one shown above, the ratio reported would be $\hat{Y}_{ratio}$ and the standard error $\text{se}(\hat{Y}_{ratio}) = \sqrt{\text{var}(\hat{Y}_{ratio})}$. Again, this estimator could prove more precise than the estimator in (6), but it requires we know *X*.

Table 3 below summarizes the useful statistical keywords available in the PROC SURVEYMEANS statement as they relate to ratios. Note that there is no coefficient of variation available like there is with the estimated total and mean, and that confidence intervals are requested using the same syntax as for means.

Keyword	Output
MEAN	Estimated mean of each variable listed in the VAR statement or estimated proportion of all levels of each categorical variable listed in the CLASS statement
STDERR	Standard error of the estimated mean, output by default when the MEAN keyword is used
VAR	Variance of the estimated mean
CV	Coefficient of variation (MEAN divided by STDERR) of the estimated mean
CLM	Confidence interval of the estimated mean based on the significance level given in the ALPHA= option (default is ALPHA=.05) and complex survey degrees of freedom (# PSUs - # strata)

Table 3. Summary of PROC SURVEYMEANS Statement Statistical Keywords Related to Estimated Ratios

QUANTILES

The fourth and final class of statistics we will discuss in this paper is *quantiles*. These became available much more recently than the first three, with the release of SAS version 9.2. Although the term "quantiles" may not be immediately familiar to every reader, the concept of them likely is. These include measures such as medians and percentiles. The sorted values of y_i are called the *order statistics*, and we will denote them $y_{(j)}$ ($y_{(1)} < y_{(2)} < \ldots < y_{(U)}$). The median is the midpoint of the order statistics. That is, the point in the order statistics sequence at which 50% of the unique values fall below it and 50% remain above.

The formal mathematical definition of quantiles requires us to first define the *cumulative density function* (CDF) for a variable *y* as

$$F(y_{(j)}) = \frac{\sum_{i=1}^{N} I(y_i \le y_{(j)})}{N} \quad (10)$$

where $I(y_i \le y_{(j)})$ is a 0/1 indicator variable of whether y_i is less than or equal to the given value $y_{(j)}$. The CDF is an increasing step function that runs from 0 to 1. The population median $y_{\gamma=.5}$ can then be defined as the smallest value of *y* such that the CDF is greater than or equal to 0.5. Equivalently, this is termed the $\gamma = 0.50$ quantile or the 50th percentile.

The CDF can be estimated from a complex survey sample data set using the weights as follows:

$$\hat{F}(y_{(j)}) = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} I(y_{hij} \le y_{(j)})}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}} \quad (11)$$

where the order statistics are based on the sample. There is no way to plot the estimated CDF from within PROC SURVEYMEANS, but one can be generated using the CDFPLOT statement in PROC UNIVARIATE. Be advised this statement will not work in conjunction with a WEIGHT statement, but the FREQ statement can be used instead. Note that the FREQ statement truncates the variable specified to the nearest integer. For large weights—in the hundreds or thousands, say—the truncation will be inconsequential, but for smaller weights it would be wise to simply multiply all weights by 100. A common weight inflation factor will have no impact on the appearance of CDF.

PROC SURVEYMEANS estimates a population quantile y_γ from the survey data set by first finding the value of j such that $\hat{F}(y_{(j)}) \le \gamma < \hat{F}(y_{(j+1)})$. If we denote this value $y_{(j)}$, the γ^{th} quantile is estimated as

$$\hat{y}_\gamma = y_{(j)} + \frac{\gamma - \hat{F}(y_{(j)})}{\hat{F}(y_{(j+1)}) - \hat{F}(y_{(j)})}(y_{(j+1)} - y_{(j)}) \tag{12}$$

The second term is an interpolation correction factor whenever $\gamma > \hat{F}(y_{(j)})$. If $\gamma = \hat{F}(y_{(j)})$, the term is zero and reduces to simply $y_{(j)}$. The only additional exceptions worth pointing out are that whenever $\gamma < \hat{F}(y_{(1)})$ or $\gamma > \hat{F}(y_{(U)})$, the quantiles default to the minimum or maximum in the sample data set, respectively.

The standard error of a given quantile is computed using a method attributable to Woodruff (1952) discussed in Dorfman and Valliant (1993) along with a few competing methods. It involves several steps. We will define them first, and then unify concepts with the help of an annotated visualization.

The method begins by computing an estimated variance of $\hat{F}(\hat{y}_\gamma)$ with respect to the complex sample design as

$$\mathrm{var}(\hat{F}(\hat{y}_\gamma)) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left(d_{hi} - \frac{\sum_{i=1}^{n_h} d_{hi}}{n_h} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \tag{13}$$

where $d_{hi} = \dfrac{\sum_{i=1}^{n_h} \sum_{i=1}^{m_{hi}} w_{hij} \left[I(y_{hij} \le \hat{y}_\gamma) - \hat{F}(\hat{y}_\gamma) \right]}{\sum_{i=1}^{n_h} \sum_{i=1}^{m_{hi}} w_{hij}}$

The estimated variance in (13) is used to form a confidence interval around $\hat{F}(\hat{y}_\gamma)$ using $t_{df,\alpha/2}$, a reference t-statistic with df complex survey degrees of freedom and significance level α. If we denote these endpoints $\hat{F}^L(\hat{y}_\gamma)$ and $\hat{F}^U(\hat{y}_\gamma)$, respectively, the next step is to invert the estimated CDF at these two points using (12). If we label the two resulting values $\hat{y}_\gamma^L$ and $\hat{y}_\gamma^U$, the next step is to solve an implied $\hat{y}_\gamma$ confidence interval equation for the standard error as follows:

$$\mathrm{se}(\hat{y}_\gamma) = \frac{\hat{y}_\gamma^U - \hat{y}_\gamma^L}{2 \times t_{df,\alpha/2}} \tag{14}$$

There is no option available to request the variance of the quantile be output, although it can easily be obtained by simply squaring the quantity in (14). As was done for the three other statistics covered in this paper, there will be a table given towards the end of this section summarizing all statistical keywords in the PROC SURVEYMEANS statement as they pertain to quantiles.

A visualization of the Woodruff method is extremely helpful, if not imperative, for grasping the sequence of steps involved. This is the aim of Figure 2 below, an annotated plot of the estimated CDF for an example continuous variable in a survey data set. The first step is to form a confidence interval around $\hat{F}(\hat{y}_\gamma)$. The second step is to translate the endpoints back to the variable scale to get quantile interval endpoints $\hat{y}_\gamma^L$ and $\hat{y}_\gamma^U$. The distance between these two endpoints provides the basis for solving for $\mathrm{se}(\hat{y}_\gamma)$ (and thus $\mathrm{var}(\hat{y}_\gamma)$). In the final step, PROC SURVEYMEANS uses $\mathrm{se}(\hat{y}_\gamma)$ to form confidence limits on the estimated quantile as $\hat{y}_\gamma \pm t_{df,\alpha/2}\mathrm{se}(\hat{y}_\gamma)$. (You can output the asymmetric confidence limits labeled (2) in Figure 2 by specifying the NONSYMCL option in the PROC SURVEYMEANS statement.)

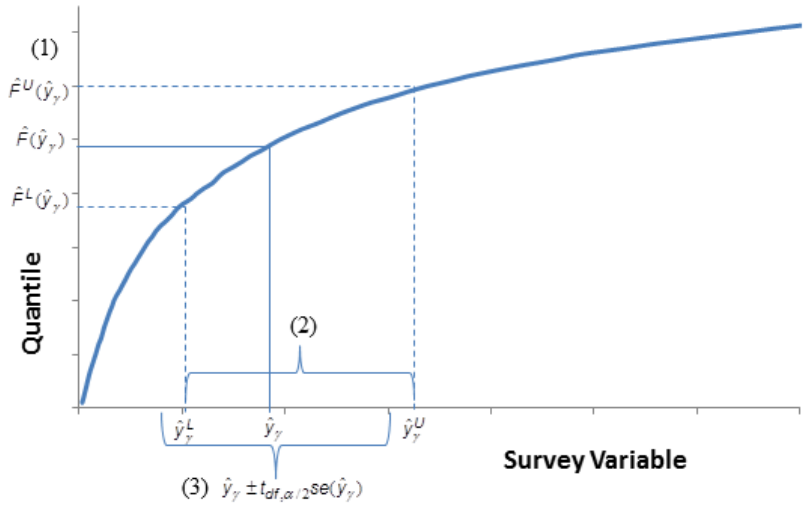

Figure 2. Visualization of the Woodruff Method for Quantile Variance Estimation.

Suppose we were interested in estimating the *quartiles* of annual electricity expenditures in U.S. dollars of all buildings in the 2003 CBECS population. Quartiles are defined as the γ = 0.25, 0.50, and 0.75 quantiles of the variable—ELEXP8, in this case. The example below demonstrates the syntax to have PROC SURVEYMEANS output these estimates as well as their standard errors and 95% confidence limits. In addition to the QUARTILES keyword in the PROC statement, MEAN is also specified to permit a comparison of the estimated median with the mean. We can observe the mean ($17,930) is significantly larger than the median ($4,301), a sign that the electricity expenditures distribution is heavily right-skewed.

```
proc surveymeans data=CBECS_2003 quartiles mean;
  strata STRATUM8;
  cluster PAIR8;
  var ELEXP8;
weight ADJWT8;
run;
```

Statistics

Variable	Mean	Std Error of Mean
ELEXP8	17930	995.072145

Quantiles

Variable	Percentile	Estimate	Std Error	95% Confidence Limits	
ELEXP8	25% Q1	1619.675091	100.746908	1416.6330	1822.7171
	50% Median	4300.740709	298.693274	3698.7640	4902.7174
	75% Q3	12664	897.759484	10855.1304	14473.7611

In the example above, note that placing the keyword QUARTILES in the PROC statement is tantamount to specifying QUANTILE=(.25 .50 .70), PERCENTILE=(25 50 75), or Q1 MEDIAN Q3. Needless to say, there are a variety of ways to request estimated quantiles. These are summarized in Table 4 alongside other pertinent statistical keywords. With the exception of the minimum and maximum value, standard errors and confidence limits are output by default any time quantiles are requested. At the time of this writing, there are no additional keywords available relating to measures of uncertainty.

Keyword	Output
MIN	Minimum value
MAX	Maximum value
RANGE	MAX – MIN
Q1	Lower quartile (25th percentile)
MEDIAN	Median (50th percentile)
Q3	Upper quartile (75th percentile)
DECILES	The 10th, 20th, …, 90th percentiles
PERCENTILES=(values)	User-defined percentiles specified as whole numbers between 0 and 100, separated by a space or comma
QUANTILES=(values)	User-defined quantiles specified as decimals between 0 and 100, separated by a space or comma

Table 4. Summary of PROC SURVEYMEANS Statement Statistical Keywords Related to Estimated Quantiles

CONCLUSION

This paper was comprised of four sections, one for each of the four major classes of statistics that can be estimated using PROC SURVEYMEANS—namely, totals, means, ratios, and quantiles. The syntax examples were intentionally rudimentary to illustrate the fundamental concepts with minimal distraction. Although it would have been excessive to demonstrate outputting all of the available statistical keywords, a tabular summarization of those most pertinent to the underlying statistic was provided at the end of each section. Note that if you ever want to simply canvass all reportable statistics associated with the underlying analysis, you can specify the ALL keyword in the PROC statement.

The reader may have also observed all examples involved estimation at the population level. *Domain analysis* is the term reserved for estimates focused only on a subset of the population (e.g., a particular region of the country or building type). It is not advised to simply subset the data for the domain of interest or use a BY statement; instead, you should use the DOMAIN statement (available in all SURVEY procedures) or create a domain-specific weight in a prior DATA step. These concepts are discussed in Lewis (2013a).

REFERENCES

Cochran, W. (1977). *Sampling Techniques*. *Third Edition*. New York: Wiley.

Dorfman, A., and Valliant, R. (1993). "Quantile Variance Estimators in Complex Surveys," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 866 – 871.

Kalton, G. (1983). *Introduction to Survey Sampling*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-035. Newbury Park, CA: Sage.

Kish, L. (1965). *Survey Sampling*. New York, NY: Wiley.

Lewis T. (2013a). "Considerations and Techniques for Analyzing Domains of Complex Survey Data," *Paper presented at the SAS Global Forum*. San Francisco, CA, April 28 – May 1. Available on-line at: http://support.sas.com/resources/papers/proceedings13/449-2013.pdf

Lewis T. (2013b). "Analyzing Categorical Variables from Complex Survey Data Using PROC SURVEYFREQ," *Paper presented at the Midwest SAS Users Group (MWSUG) Conference*. Columbus, OH, September 22 – 24.

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Mukhopadhyay, P., An, A., Tobias, R., and Watts, D. (2008). "Try, Try Again: Replication-Based Variance Estimation Methods for Survey Data Analysis in SAS® 9.2," *Paper presented at the SAS Global Forum*. San Antonio, TX, March 16 – 19. Available on-line at: http://www2.sas.com/proceedings/forum2008/367-2008.pdf

Woodruff, R. S. (1952). "Confidence Intervals for Medians and other Position Measures," *Journal of the American Statistical Association*, **47**, pp. 635 – 646.

CONTACT INFORMATION

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

Taylor Lewis
Joint Program in Survey Methodology (JPSM)
1218 LeFrak Hall
University of Maryland

College Park, MD 20742
Email: tlewis@survey.umd.edu