Paper 266-2010

Principles of Proper Inferences from Complex Survey Data

Taylor Lewis, University of Maryland, College Park, MD

ABSTRACT

This paper discusses commonly-encountered features of complex survey datasets that should be accounted for by employing the SURVEY family of SAS® procedures. Concepts of clustering, stratification, unequal weighting, finite population corrections, and replication variances approximation methods new to SAS 9.2 will be presented. Potential pitfalls will be highlighted as will a few methods to trick SAS into performing techniques not yet available, such as DOMAIN analysis under replication variance approximations.

FEATURES SPECIFIC TO COMPLEX SAMPLES

A complex sample survey is defined as any sample selection scheme other than simple random sampling (SRS), the more familiar scheme in which every potentially sampled element has an identical and known, non-zero probability of selection. Data collected from complex sample surveys requires special attention due to four commonly present features: clustering, stratification, unequal weights, and finite populations. Whereas we might use PROC MEANS to report averages and standard errors from a physical science experiment, surveys of populations (especially humans) often require employing PROC SURVEYMEANS to properly account for standard error formula modifications attributable to these design features. By walking through a hypothetical complex sample of students selected to participate in a statewide mathematics aptitude test, the necessary SAS syntax amendments will be presented.

In this paper we assume a state board of education aims to measure mathematics aptitude for 8th graders attending public schools by posing 100 questions of varying quantitative topics. A student's score is treated as a continuous variable, ranging between 0 and 100. It was determined infeasible to collect data via a simple random sample, since no universal list of students exists. There was, however, a list of all public schools within the state and corresponding geographic information. From this list, a complex sample was designed.

CLUSTERING

Target sample units sometimes naturally reside in groups, or *clusters*. From a data collection standpoint, it might be easier to identify and sample these clusters with known probability than it would be to directly sample units therein. Such is the present case where a list of schools, the clusters in which students reside, is the only sample frame readily attainable. Another common reason for clustering is to minimize travel costs to collect data—many nationwide face-to-face surveys sample clusters of counties or some other geographic unit.

Clustering is rarely ideal, as it typically decreases precision. Units within clusters often yield similar values for survey variables measured. For instance, households within the same block tend to have similar income and many other socioeconomic characteristics. Intuitively, there is less information to be gained surveying more similar people within the same cluster, at least compared to a more diversified SRS. Nonetheless, clustering is often an unavoidable feature of a large survey.

There may also be clusters sampled within clusters. In our example, the primary sampling unit (PSU) is a public school, but not all students participated. Instead, we sampled two homerooms within a school and ultimately two students within each homeroom. Despite this multi-layered nature, as a SAS programmer you only need to be concerned with identifying the PSU—more on this later.

STRATIFICATION

Stratification is a technique whereby the population is partitioned into non-overlapping groups known as *strata* (singular: *stratum*), and an independent sample is selected within each. In direct contrast to clusters, homogeneity within strata leads to greater precision for population-level estimates. Stratifying can also be employed to achieve minimum sample sizes for less prevalent groups whose precision from an SRS would be otherwise unacceptable. For instance, in a survey of a large organization, if inferences of accountants' attitudes were desired, sequestering this group of employees into their own stratum prior to sampling could ensure a large enough respondent pool from which to infer about all accountants in the organization.

In our math aptitude example, two strata were formed—urban and rural—utilizing geographical indicators from the school list. It was believed there was a substantial difference in aptitude between the two locations. If true, stratifying on this factor would reduce the overall variability of statewide estimates.

UNEQUAL WEIGHTING

Recall two schools were selected within each stratum, with subsequent sampling of homerooms and then students. Therefore, it seems unlikely students share equal selection probabilities. As a rule, when selection probabilities differ, analysis should include respondent-level weights inversely proportional to the selection probability. If a respondent has a weight of 5, say, the interpretation is that the respondent represents himself/herself and 4 other individuals in the population. Neglecting to do so may introduce bias, particularly if the weights are correlated with the outcome variable(s).

Another frequent contributing factor to unequal weighting is nonresponse. A survey may begin with equal selection probabilities, but varying patterns of nonresponse give rise to weight adjustment methods attempting to make respondents more in line with some population benchmark(s). The resulting weights, therefore, may no longer be equal.

FINITE POPULATIONS

A large class of statistical analyses assumes SRS, perhaps from an infinite or otherwise indefinable population. It is then safe to assume sampling *with replacement*, where each selection is independent of another and the covariance between any two population units is zero. When surveys collect data from a sizable portion of the population, this assumption is no longer valid. For instance, assume we survey 20 employees in a division of 100. If the sample of 20 was selected without replacement, the covariance between any two units is non-zero (and negative), and actually leads to an augmented variance formula including a *finite population correction*, or FPC. The new multiplying factor is 1 minus the sampling fraction. Hence, the variance of our survey of employees would actually be (1 – 20/100) = 0.8 times that reported from PROC MEANS. In essence, by collecting data on more of the population we are rewarded with a reduction in variance.

PUTTING IT ALL TOGETHER

Table 1 summarizes the four complex sample design issues and the corresponding statements or options required. The survey-specific syntax is identical across all four currently available SAS/STAT® analysis procedures—SURVEYMEANS, SURVEYREG, SURVEYFREQ, and SURVEYLOGISTIC—although this paper provides examples solely from PROC SURVEYMEANS.

Table 1. Summary of Complex Survey Features.

Feature	Syntax	Comments
1. Stratification	STRATA statement	Survey data set should have a variable identifying the stratum to which the observation belongs.
2. Clustering	CLUSTER statement	Survey data set should have a variable identifying the cluster to which the observation belongs. If multiple steps of cluster sampling occur, provide only identifier(s) for the primary sampling stage.
3. Unequal Weighting	WEIGHT statement	Only observations with weight > 0 are maintained. Without a WEIGHT statement, weights are assumed to equal 1.0.
4. Finite Population Correction (FPC)	TOTAL=<data-set-name> or RATE=<data-set-name> both options in the PROC SURVEYMEANS statement	Given a supplementary dataset with a like-named stratum variable and corresponding population total, _TOTAL_, SAS can calculate the stratum-specific FPC based on the number of observations present in the analysis data set. Alternatively, users can do the calculations themselves and provide via the RATE= option. _RATE_ is then the key dataset variable, a value between 0 and 1.

Note that when both stratification and clustering are involved, clusters are assumed situated within strata, but either statement can stand alone. For instance, a single-stage sample of clusters would require the CLUSTER statement but not the STRATA statement, and vice versa for a series of samples within strata, given non-clustered population elements (or, equivalently, clusters all of size one). Lastly, be advised the FPC applies only to the first-stage of sampling within strata[1].

ACCOUNTING FOR COMPLEX SURVEY FEATURES IN OUR MATH APTITUDE SURVEY

Sample code to read in the mock data set of 16 observations is as follows (2 students within 2 homerooms within 2 schools within 2 strata = 2 x 2 x 2 x 2 = 16):

```
data test;
  input stratum school homeroom weight grade tutor $;
* stratum=1 --> urban
  stratum=2 --> rural
;
datalines;
1 1 1 22   87 Y
1 1 1 22   89 N
1 1 2 18   91 Y
1 1 2 18   84 Y
1 2 1 17.5 82 N
1 2 1 17.5 94 N
1 2 2 24   93 N
1 2 2 24   94 Y
2 1 1 9    78 N
2 1 1 9    84 N
2 1 2 4.5  90 N
2 1 2 4.5  82 N
2 2 1 6    88 N
2 2 1 6    93 Y
2 2 2 8    77 Y
2 2 2 8    81 N
;
run;
```

To fully account for the clustering, stratification, and unequal weighting present in our survey, the proper syntax to estimate the state-level mean math aptitude and its standard error would be as follows:

```
proc surveymeans data=test mean stderr;
  strata stratum;
  cluster school;
  var grade;
weight weight;
run;
```

The STRATUM variable is supplied in the STRATA statement identifying an urban/rural indicator variable, while the CLUSTER statement's SCHOOL variable identifies the PSUs within each stratum. The WEIGHT statement points to the variable WEIGHT in the data set representing the inverse of the selection probability. The VAR statement requests an analysis of numeric variable GRADE, each student's math test score. Note that if you provide a character variable or include the numeric VAR variable in a CLASS statement, SAS will estimate the proportions of each variable level.

One may wonder why SAS only needs information on the primary cluster selected, when there are subsequent sampling stages. The reason is that major computational simplifications can be achieved if with replacement selection is assumed. Granted, most designs are without replacement, but with replacement is often plausible when the first-stage sampling rate is low. For instance, if the number of urban schools was 85 and only 2 were selected, assuming with replacement sampling would not be far-fetched. Moreover, assuming with replacement generally leads to overestimation of standard errors, a tolerable downside to the less involved calculations.

This issue deserves attention since one might initially be tempted to supplement the CLUSTER statement with the HOMEROOM variable also in the data set, but doing so is incorrect. If PROC SURVEYMEANS sees two variables in

[1] The FPC is calculated overall or by stratum, if applicable, based on the number of PSUs with a positive weight. Remain cognizant of item nonresponse from respondents who answered enough questions to be considered a respondent, thus having positive weights. In this instance, the sampling fraction calculated behind the scenes may be higher than it technically should be.

the CLUSTER statement, the combination of the two is assumed to identify a PSU. This would mean SAS observes four clusters per strata instead of two, leading to a significantly underestimated standard error. We can tolerate overstating measures of uncertainty, not understating.

ANALYZING DOMAINS OF DATA

By default SAS outputs population-level statistics and standard errors, yet we often wish to compute estimates for domains—also termed *subdomains* or *subgroups*. Under most circumstances we need to employ the DOMAIN statement to do so instead of subsetting the data set or utilizing a BY statement. The key criterion to consider is whether the sample design is based on the domain of interest. For instance, if we wanted mean exam grade estimates for the rural stratum, subsetting the data set would be appropriate, as would putting the STRATUM variable in a BY statement. Because we sampled based on this factor, the domain sample sizes are fixed.

Consider another scenario where we want to compare mean exam grade for tutored students. The variable TUTOR has values Y/N indicating whether a student received tutoring prior to taking the math aptitude test. For the given analysis, the DOMAIN statement should be used, reasoning we did not sample based on whether students were tutored. In essence, the number of sampled elements from that domain is a random variable, and we need to account for the extra variability that introduces.

The following illustrates the difference between the two methods for analyzing mean grade across the TUTOR variable domain. The first PROC analyzes a subsetted data set, whereas the second properly accounts for the domain of tutored students:

```
proc surveymeans data=test (where=(tutor='Y')) mean stderr t df;
  strata stratum;
  cluster school;
  var grade;
weight weight;
run;
proc surveymeans data=test mean stderr t df;
  strata stratum;
  cluster school;
  var grade;
weight weight;
domain tutor;
run;
```

The means are the same, but the standard errors differ. Unfortunately, there is no universal expectation as to the direction with which inferences will be incorrect, because there are numerous, competing factors behind the scenes.

The first major factor is the two-PSU requirement to compute a stratum variance estimate. The rural stratum (STRATUM=2) contains only one school (i.e., PSU) with a tutored student. The DOMAIN statement properly accounts for this, but when subsetting SAS issues the following to the log: “Only one cluster in a stratum for variable(s) grade. The estimate of variance for grade will omit this stratum.” The overall variance estimate for a domain crossing multiple strata involves a summation of individual stratum-specific domain variance estimates. Therefore, eliminating the contribution of a particular stratum would necessarily underestimate the overall variance.

An erroneous degrees of freedom computation is another unfortunate byproduct of subsetting. For Taylor series variance approximations, SAS computes degrees of freedom as the number of PSUs less the number of strata. Hence, lopping off PSUs will unnecessarily reduce degrees of freedom for t-tests, widening confidence intervals requested directly from PROC SURVEYMEANS.

Lastly, the FPC (if applicable) will ignore observations not meeting the WHERE condition when calculating the sampling fraction; all else held constant, this will inflate standard errors. The sampling fraction in the FPC formula should include all PSUs regardless of whether they contain an element within the domain of interest. Again, the DOMAIN statement will handle this properly.

ANOTHER DOMAIN SCENARIO

The concept of domain analysis may surface when sampling from an imperfect list introduces out-of-scope, or *ineligible*, units that cannot be identified and removed beforehand. When developing the final respondent weights, we reason that if a portion of the sample was proven ineligible, then it is likely a portion of those not sampled is also ineligible. For this reason, records of ineligible units are sometimes maintained alongside the respondents in the analysis data set with positive weights. This may seem confusing, but you do not want to exclude them prior to running PROC SURVEYMEANS. Instead, create an indicator variable based on eligibility and retain results only for the eligible domain.

Returning to our math aptitude example, assume the target population was 8th graders who are also U.S. citizens. During the survey administration, suppose we determined a particular student is a foreign exchange student. We will assume such is the case with the first student sampled from the first homeroom in the urban stratum. Instead of excluding this individual from the data set TEST, we would create a domain indicator variable as follows, focusing only on the portion of output where ELIG='Y':

```
data test;
  set test;
if stratum=1 and school=1 and homeroom=1
  then elig='N';
else elig='Y';
run;
proc surveymeans data=test mean stderr;
  stratum stratum;
  cluster school;
  var grade;
weight weight;
domain elig;
run;
```

Be aware that oftentimes when ineligibility is determined, data collection terminates, leaving analysis variables missing for ineligible cases. But observations with missing values for VAR variables are excluded from analysis[2]. This is an undesirable consequence for the given scenario. To mitigate, simply fill in some value for the analysis variables of ineligible cases. These arbitrary values will be harmless, since they are computed within the ineligible domain you do not care about, but incorrect results can occur if the values are left missing.

WHEN THE DOMAIN STATEMENT IS UNAVAILABLE

Without getting bogged down in computational details, a quick aside regarding how SAS develops domain standard errors is useful to derive a method for producing standard errors when a DOMAIN statement is not allowed—currently the case with the RATIO statement.

For the domain defined by TUTOR='Y', SAS creates a new weight variable equaling the original weight for observations meeting that condition. Recall that if no WEIGHT statement is provided, all weights are assumed to be 1.0. Where TUTOR not equal 'Y' (and TUTOR not missing), the new weight variable is set to zero. SAS then proceeds to compute a mean and standard error for variable(s) listed in the VAR statement with the newly created weight variable in standard fashion according to sample design. This implies we could create a new weight equaling the original weight for those with TUTOR='Y' and zero otherwise, insert this new variable in the WEIGHT statement and match the DOMAIN statement's output. This will not work, however, since observations with non-positive weights are excluded before performing any computations. The go-around, then, is to create a new weight variable equaling some minuscule value strictly greater than 0, for example 0.00000001.

REPLICATION VARIANCE METHODS

Thus far, we have been utilizing the default variance estimation strategy in SAS' SURVEY family of procedures, Taylor series linearization. (Recall the standard error, our measure of uncertainty discussed up until this point, is merely the square root of the variance.) In the PROC statement, the Taylor series method can be explicitly requested with the option VARMETHOD=TAYLOR. There exist alternative strategies, however, for estimating variances based on replication, or repeated subsampling of the full survey data set. If set up properly, computing a variance for a specific estimate like a mean can be approximated by comparing variability among subsamples. These methods are new to SAS in Version 9.2, and although this paper will touch on many of the basic principles, a more thorough discussion is given in Mukhopadhyay et al. (2008).

The math aptitude example purposefully illustrated selecting two PSUs per stratum, since this is a very common sample design, one SAS programmers will frequently encounter when analyzing public-use survey data sets. Such a method allows for maximum stratification while still allowing variances within strata to be computed. The first replication method, *balanced repeated replication*, is geared specifically for this design.

[2] By specifying the MISSING option in the PROC statement, SURVEYMEANS will treat missing values as a legitimate category, but this is only applicable to missing categorical analysis variables and missing stratum, cluster, or domain variables.

BALANCED REPEATED REPLICATION

Balance repeated replication (BRR) is a subsampling scheme that creates *R* replicates by selecting one of the two PSUs from each stratum. Without delving too deep into the theory, the number of replicates will always be the next multiple of four strictly greater than the number of strata, *H*. That is, $H < R \leq H + 4$. And each replicate is defined by a new weight variable, collectively termed *replicate weights*. The weight of the selected PSU is doubled and the unselected PSU is set to zero. With the new set of replicate weights, the estimate is repeatedly calculated and variance computed as $\frac{1}{R}\sum_{i=1}^{R}\left(\hat{\theta}_r - \hat{\theta}\right)^2$, where $\hat{\theta}$ is the estimate from the *r*-th replicate weight variable and $\hat{\theta}$ is the full sample estimate based on the original weight. SAS users can request BRR by supplying the VARMETHOD=BRR option in the PROC SURVEYMEANS statement.

FAY'S METHOD

Fay's method is a variant of BRR where instead of doubling one PSU's weight and setting the other to 0, you might set one equal to 1.5 times the original and the other 0.5, such that each PSU is included in all replicate estimates. To employ Fay's method, add the option FAY<=decimal> in parenthesis after VARMETHOD=BRR. The decimal is optional; without specifying a value, the default is 0.5. If we denote the decimal as ε, the modified variance formula becomes $\frac{1}{R(1-\varepsilon)^2}\sum_{i=1}^{R}\left(\hat{\theta}_r - \hat{\theta}\right)^2$.

JACKKNIFE REPEATED REPLICATION

Another viable replication technique is *jackknife repeated replication* (JRR) in which you delete one PSU at a time and calculate the estimate with the remaining PSUs. Similar to BRR, this method produces replicate weights set to (i) zero for all observations in the omitted PSU (ii) the original weight multiplied by a factor of $\frac{n_h}{n_h - 1}$, where n_h equals the number of PSUs from the *h*-th stratum[3], for observations within another PSU in the stratum containing the omitted PSU or (iii) the unmodified weight if within a PSU outside the omitted PSU's stratum. JRR is more flexible than BRR, able to accommodate any number of PSUs per stratum. Note, however, this method produces equal number replicates as PSUs in the data set, which could be significantly greater than the number of replicates obtained from BRR if there are strata with many more than 2 PSUs.

The JRR variance formula is $\sum_{i=1}^{R}\frac{n_h - 1}{n_h}\left(\hat{\theta}_r - \hat{\theta}\right)^2$, where $\hat{\theta}_r$ and $\hat{\theta}$ are defined as before using the *r*-th replicate weight estimate or original weight estimate, respectively, and the factor $\frac{n_h - 1}{n_h}$ is called the *jackknife coefficient*. Notice how this coefficient is applied to the summed, squared deviations in lieu of the replicates' deviations being averaged over all *R* replicates as with BRR.

SYNTAX EXAMPLES

There are two paths to analyzing survey data via replication methods. One begins with a full sample weight and constructs replicate weights on the fly for each invocation of PROC SURVEYMEANS. The second path calls upon a survey data set already containing the replicate weights. The second path is recommended since it is computationally and syntactically more efficient. SAS has an OUTWEIGHTS=<data-set-name> option available after the VARMETHOD=<method> option to quickly realign a programmer onto the second path if initially only provided a file containing only one weight variable.

The next line of SAS code demonstrates how to analyze our two-PSU-per-stratum design via BRR and then store the BRR replicate weights in a new data set for later use:

[3] If the design is not stratified, the n_h simplifies to *n*, the overall number of PSUs in the sample.

```
proc surveymeans data=test varmethod=BRR (outweights=test_brr);
  strata stratum;
  cluster school;
  var grade;
weight weight;
run;
```

The listing now displays a BRR-derived standard error approximation, and the new analysis data set TEST_BRR contains all original TEST variables in addition to four replicate weights (again, the number of BRR replicates is the first multiple of four greater than the number of strata) named RepWT_1-RepWT_4.

Henceforth, to analyze the survey data one points PROC SURVEYMEANS to TEST_BRR and the replicate weights with the following simplified syntax. One can verify standard errors generated from syntax below will match precisely output generated from syntax above, only this time SAS had no need to recalculate the BRR weights.

```
proc surveymeans data=test_brr varmethod=BRR;
  var grade;
weight weight;
repweights RepWt_1-RepWt_4;
run;
```

There are several points to observe. For one, the STRATA and CLUSTER statements are no longer needed, since the variance is based solely upon variation between the replicates' mean estimates. If they were provided, SAS would simply ignore them. From the perspective of data confidentiality, this is a nice feature of replication methods. Strata and clusters often denote geographic locations, so omitting them from a public-use data set helps prevent disclosure. Another key point is that the original weight is still provided and is used for the overall mean estimate, $\hat{\theta}$. If only the REPWEIGHTS statement is provided, the original weight is interpolated as the average of replicate weights for that observation, which may not perfectly match a similar run in which the WEIGHT statement is included. Results would still be close, however, so observing majorly disparate variances is evidence some additional syntax or set-up errors may have occurred.

Many of the BRR syntax adjustments will work for VARMETHOD=JACKKNIFE|JK, but the jackknife coefficients add another dimension. The option OUTJKCOEFS=<data-set-name> under the jackknife replication method will create an output data set with three key variables:

- Replicate – the replicate number
- JKCoefficient – the factor $\frac{n_h - 1}{n_h}$ corresponding to the omitted PSU's stratum
- DonorStratum – indicator of stratum from which the PSU was deleted (only applicable with stratification)

The following illustrates how the OUTWEIGHTS=<data-set-name> option is requested as before, but the OUTJKCOEFS=<data-set-name> option stores jackknife coefficients as a variable named for each replicate in a separate supplementary data set.

```
proc surveymeans data=test varmethod=JK (outweights=test_JK
outjkcoefs=test_JK_coef);
  strata stratum;
  cluster school;
  var grade;
weight weight;
run;
```

To draw upon the data set TEST_JK and its replicate weights, utilize the REPWEIGHTS statement as before, but now include the JKCOEFS=<data set-name> option in the same statement pointing PROC SURVEYMEANS to the necessary jackknife coefficients. Technically, you can provide individual coefficients corresponding to the *r*-th replicate, but there seems less room for error letting SAS do the work. Once again, the STRATA and CLUSTER statements are no longer needed.

```
proc surveymeans data=test_jk varmethod=JK;
  var grade;
weight weight;
repweights RepWt_1-RepWt_4 / jkcoefs=test_JK_coef;
run;
```

Warnings:

- Be sure to match VARMETHOD with proper replicate weights. In the math aptitude test example, specifying the data set TEST_JK but requesting VARMETHOD=BRR or specifying TEST_BRR but requesting VARMETHOD=JK will produce no warning or error in the log, yet results would be erroneous.
- Be sure to include all replicate weights in the REPWEIGHTS statement. In the above example, SAS will appear to run as normal even if the statement incorrectly called on RepWt_1 - Rep_Wt_2.
- If you provide a REPWEIGHTS statement without declaring VARMETHOD, the default is VARMETHOD=JK.
- If you fail to provide a supplementary data set with the jackknife replicate coefficients, the default value for all replicates is $\frac{R-1}{R}$, where *R* is determined from the number of replicate weights.

FINAL COMMENTS ON REPLICATION METHODS

All variance approximation strategies mentioned in this paper tend to produce similar results with reasonably large sample sizes. While the Taylor series linearization approach is arguably the most common variance approximation method employed, there are potential advantages to pursuing a replication variance strategy. Previously discussed was the fact that replicate weights contain all design information necessary, such that (possibly sensitive) stratum and cluster identifiers need not be released on a micro-data file. Another advantage is that replication can better account for variability attributable to nonresponse weighting adjustments if the adjustments are performed separately for each replicate weight set. Additionally, this paper's emphasis on the need for a DOMAIN statement when analyzing subgroups does not apply to replication techniques. SAS will send an error to the log when attempting to use a DOMAIN statement with a VARMETHOD other than Taylor series, but subsetting is justifiable (Brick et al. 2000, p. 7)[4].

Lastly, the TOTAL=<data-set-name> option is unavailable for replication methods. Thus, stratum-specific FPCs are only allowed under Taylor series linearization. This restriction is probably wise, since the theory behind replication variances is based on with-replacement sampling, or without-replacement sampling at a low sampling rate, such that it could be safely treated as with-replacement. Applying an FPC seems inappropriate and, moreover, would only marginally reduce variances. Although there is technically an FPC that could be incorporated as part of the jackknife coefficients under JRR, there is no real FPC applicable to a BRR replicate, since PSUs are taken from different strata with potentially different sampling rates.

CONCLUSION

When analyzing data collected from a complex sample survey, SAS programmers should be aware of several distinctive features: stratification, clustering, unequal weights, and (less frequently) finite population corrections. When such features are evident, the SURVEY family of PROCs must be employed to make proper inferences. Most non-SURVEY procedures allow for a WEIGHT statement, which is all that would be needed to create unbiased estimates of means or other parameters such as regression coefficients, yet the standard errors would not be correct since with-replacement sampling sans clusters and strata is implicitly assumed.

This paper outlined syntax to properly incorporate these survey design features into one's analysis. Concepts and considerations of replication variance strategies, now available in SAS/STAT as part of Version 9.2, were also mentioned. Where deemed necessary, potential pitfalls were noted and recommendations for getting around as-yet unavailable analyses were offered. Hopefully, the reader now boasts a better understanding of these characteristics pertinent to survey data and has improved his or her confidence in carrying out sound analyses.

REFERENCES

Brick, M., Morganstein, D., and Valliant, R. (2000). "Analysis of Complex Sample Data Using Replication." Westat Technical Paper. Available at http://www.westat.com/wesvar/techpapers/ACS-Replication.pdf

Kalton, G. (1983). *Introduction to Survey Sampling*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-035, Beverly Hills and London: Sage Publications.

[4] This was discovered while researching for the paper. As the abstract hinted, I had derived a method similar to what was suggested for domain estimation with a RATIO statement, setting any and all replicate weights equal to some miniscule value. The go-around is not necessary, however, and was excluded.

Kreuter, F., and Valliant, R. (2001). “A Survey on Survey Statistics: What is Done and Can be Done in Stata.” *The Stata Journal*. 1(1), pp 1-22.

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 2008 SAS Global Forum, March 16 – 19, San Antonio Texas. Available at http://www2.sas.com/proceedings/forum2008/367-2008.pdf

SAS Institute Inc. (2008). SAS/STAT® 9.2 User’s Guide. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Name: Taylor Lewis
Enterprise: Joint Program in Survey Methodology, University of Maryland
Address: 1218 LeFrek Hall
City, State ZIP: College Park, MD 20742
E-mail: tlewis@survey.umd.edu