Paper 767-2017

Estimation Strategies Involving Pooled Survey Data

Taylor Lewis, George Mason University

ABSTRACT

Pooling two or more cross-sectional survey data sets (i.e., stacking the data sets on top of one another) is a strategy often utilized by researchers for one of two purposes: (1) to more efficiently conduct significance tests on point estimate changes observed over time; or (2) to increase the sample size in hopes of improving the precision of a point estimate. The latter purpose is especially common when making inferences on a subgroup, or domain, of the target population insufficiently represented by a single survey data set. The aim of this paper is to walk through a series of practical estimation objectives that can be tackled by analyzing data from two or more pooled survey data sets. Where applicable, the resulting interpretive nuances are discussed.

1. INTRODUCTION

The aim of this paper is to discuss strategies to formulate point estimates and measures of variability for pooled survey data sets. By "pooled," we mean two or more data sets combined (i.e., stacked) into one. This is especially common in repeated cross-sectional survey efforts (e.g., Lee et al., 2007), although there are other contexts where it could be applicable as well. The paper is structured as follows. Section 1 provides a brief background on the features of complex survey data. Section 2 introduces a training data set used in examples appearing throughout the paper. Section 3 covers strategies to conduct significance tests on point estimate changes observed across two or more pooled survey data sets, and Section 4 discusses considerations to formulate a single point estimate from the combined data sets, whereas Section 6 touches on considerations for pooling data to form a single point estimate. Finally, Section 6 concludes with a summary of the key takeaway points.

2. BACKGROUND ON FEATURES OF COMPLEX SURVEY DATA

This section provides a background on the definition of "complex" survey data—for an even more in-depth discussion, see Chapter 1 of Lewis (2016). In particular, contingent on the sample design utilized, complex survey data may have one or more of the following four features:

- Finite population corrections
- Clustering
- Stratification
- Unequal weights

In general, when data emanate from a sample design that involved one or more of these features, it is recommended that you use a SAS/STAT® analysis procedure prefixed by SURVEY, as this family of SAS® procedures shares a common syntax structure that can be used to identify any of these features in the input data set so that point estimates and measures of variability can be computed properly.

In many introductory statistics courses, the implied data collection mechanism is simple random sampling with replacement, 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. Indeed, it seems simple random sampling is the exception rather than the rule.

For sake of an example, suppose a high school's administrative staff seeks to measure the mathematical aptitude of its N = 1,000 students by way of a standardized test. To be specific, the finite population of interest is the student body of the high 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 traditional statistical theory that the sample mean

$$\hat{\bar{y}} = \frac{\sum_{i=1}^{n} y_i}{n}$$

is an unbiased estimate of the true population mean

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

or the average test score for all students in the high school. If the sample were selected with replacement, meaning each student could be sampled (and measured via the test) more than once, then the estimated variance of the sample mean would be calculated as

$$\text{var}(\hat{\bar{y}}) = \frac{1}{n} \frac{\sum_{i=1}^{n} \left(y_i - \hat{\bar{y}}\right)^2}{(n-1)}$$

If students were selected without replacement, however, the variance formula would be modified to

$$\text{var}(\hat{\bar{y}}) = \frac{1}{n} \frac{\sum_{i=1}^{n} \left(y_i - \hat{\bar{y}}\right)^2}{(n-1)} \left(1 - \frac{n}{N}\right)$$

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 only that of the sample mean. Notice that as the sampling fraction 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. This makes sense, because the estimate does not vary from one sample to the next; it is always equal to the population parameter.

One way to incorporate the FPC is to use the TOTAL= option in the PROC statement. SAS determines the sample size, n, from the input data set, but relies on the user to specify the population size, 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 that may be present in a complex survey data set 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. From the standpoint of data collection logistics, it would be much easier to sample homerooms and administer the test therein than to track down each sampled student independently. Whenever this is done, however, the clustering should be accounted for during the analysis stage by specifying the cluster identifier variable (e.g., homeroom) in the CLUSTER statement of the respective SURVEY procedure.

There is no mandate to sample all units within a cluster. For example, we could select two students within each of a sample of homerooms. It should be emphasized, however, that only the primary sampling unit (PSU) identifier should be specified in the CLUSTER statement—in this case, the homeroom code. 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 (Kalton, 1983) frequently adopted in practice to simplify variance estimation calculations. A commonly voiced concern is that this will not account for all stages of sampling and, thus, may underestimate variability. More commonly, however, the result is a slight overestimation of variability—for more discussion on this point and a few references to empirical investigations into the matter, see Section 1.4.4 of Lewis (2016).

The third feature of complex survey data is stratification, which arises when all 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, except in rare circumstances, stratification increases precision. The reason is that the overall variance is a function of the sum of stratum-specific variances. 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 the high school mathematics aptitude example, a prudent stratification variable might be grade level. Suppose the homerooms could be grouped into four 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.

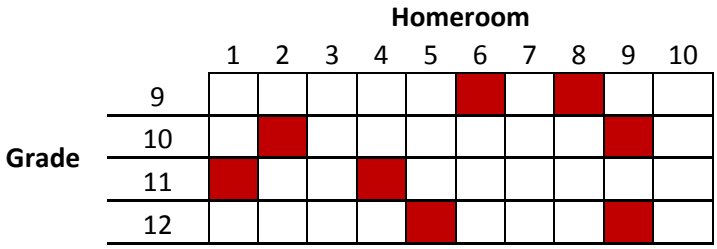

Figure 1. Visual Representation of a Stratified, Cluster Sample for the Example High School Mathematics Aptitude Survey

In general, whenever sampling rates vary amongst the ultimate sampling units, one should account for this by assigning a unit-level weight equaling the inverse of that unit's selection probability. Unequal weights are the fourth feature of complex survey data and can be interpreted as the number of population units a sampling unit represents. For instance, if a sampling 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. The 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.

Note that a complex survey data set can have any combination of the four features discussed in this section. Perhaps the most frequent combination is stratification, clustering and unequal weights. To name but a few examples, see the user documentation for the public release files of the National Health Interview Survey, the National Survey of Family Growth, the National Ambulatory Care Survey, the National Immunization Survey, or the Commercial Buildings Energy Consumption Survey.

3. EXAMPLE SURVEY DATA

To help facilitate exposition of the strategies presented in this paper, let us introduce an example survey whose sample design produces an analysis data set with three features of complex survey data: clustering, stratification, and unequal weights. Following similarly in spirit the discussion in Section 2, suppose a high school (grades 9 through 12) administers a mathematics aptitude survey to a sample of students. Suppose further that the students are all assigned to one homeroom in which they begin the first 15 minutes of each school day for purposes of an attendance check and other administrative tasks. For ease of test administration, the sample design calls for the homerooms to be stratified by grade level (CLASS), and then two homerooms are selected within each grade (HOMEROOM=1 or HOMEROOM=2), and ultimately two students are selected at random within each homeroom to take a maximum 100-point test. Hence, a total of 16 students are sampled. One version of the test is administered at the beginning of the school year (GRADE1) and another version is administered on the same sampled students at the end of the school year (GRADE2), such that a more precise estimate of progress over the course of the school year can be assessed. The variable WEIGHT compensates for the unequal selection probabilities of 16 sampled students.

Furthermore, assume that this is an ongoing effort, with an independent sample of students taken each school year. Data from the 2016 survey is stored in the data set GRADES_2016, and data from 2017 survey is stored in the data set GRADES_2017. After reading in these two survey data sets, the DATA step at the bottom of the syntax below stacks the two cross-sectional survey data sets on top of one another into a single data set called GRADES with a numeric indicator variable YEAR assigned as 0 if the record was from the 2016 data set and as 1 if from the 2017 data set. Notice how the PSU codes for the 2017 survey administration are modified to new codes, but the stratum codes are retained. Without this adjustment, because the PSUs are coded arbitrarily as 1 and 2, the SURVEY procedure would treat cases in 2016 and 2017 as emanating from the same homerooms within grade level, which would not be appropriate. A grade level code of 9 means the same thing in 2016 as it does in 2017, but a homeroom code of 1 does not, because it is comprised of a completely fresh crop of students.

```
*** read in and combine mathematics aptitude survey data sets;
data grades_2016;
  input class class_type $ homeroom grade1 tutor $ grade2 weight;
datalines;
12 upper 1 87 Y 94 49.5
12 upper 1 89 N 89 49.5
12 upper 2 91 Y 94 48
12 upper 2 84 N 92 48
11 upper 1 82 N 84 47.5
11 upper 1 94 N 95 47.5
11 upper 2 93 N 95 48
11 upper 2 94 Y 97 48
10 under 1 78 N 81 39
10 under 1 84 N 84 39
10 under 2 90 N 87 37.5
10 under 2 82 N 85 37.5
 9 under 1 88 N 88 40
 9 under 1 83 Y 89 40
 9 under 2 77 N 85 48
 9 under 2 81 N 84 48
;
run;
```

```
data grades_2017;
  input class class_type $ homeroom grade1 tutor $ grade2 weight;
datalines;
12 upper 1 88 Y 95 51.5
12 upper 1 91 N 92 51.5
12 upper 2 87 Y 92 49
12 upper 2 87 N 91 49
11 upper 1 83 N 87 47.5
11 upper 1 93 Y 98 47.5
11 upper 2 94 Y 97 47
11 upper 2 95 Y 99 47
10 under 1 81 N 84 37
10 under 1 83 N 84 37
10 under 2 92 Y 88 39.5
10 under 2 81 N 86 39.5
 9 under 1 89 N 91 41
 9 under 1 92 Y 92 41
 9 under 2 81 N 87 46.5
 9 under 2 79 N 87 46.5
;
run;

* pool data sets;
data grades;
  set grades_2016 (in=a)
      grades_2017 (in=b);
if a then year=0;
  else year=1;

* create new homeroom (i.e., PSU) codes for second year;
if year=1 then homeroom=homeroom+2;
run;
```

Note that, for the sake of data collection efficiency and to improve the precision of estimates of change over time, survey designers occasionally maintain the same strata/PSU structure, even though a "fresh" sample of subsequent sampling stage units is selected within PSUs across years. This is especially common in nationally representative surveys with in-person interviewing. In those instances, you would not want to perturb the PSU codes, as the survey designers likely maintained the PSU coding structure purposefully. If applicable, this should be noted in the data user documentation and/or example syntax files released by the survey administration team in tandem with the raw data.

4. CONDUCTING SIGNIFICANCE TESTS ON POINT ESTIMATE CHANGES OVER TIME

4.1 INTRODUCTION

In this section, we present strategies to conduct significance tests on point estimate changes over time. In terms of traditional statistical hypothesis testing notation, this is to say we wish to test

$$H_0\text{: } \bar{d} = \theta_2 - \theta_1 = 0 \text{ vs. } H_1\text{: } \bar{d} = \theta_2 - \theta_1 \neq 0$$

(although tests can be one-sided as well), where the first time period is subscripted with a 1 and the second time period is subscripted with a 2. We use the general "theta" notation here to emphasize the generalizability to any finite population parameter.

To carry out the hypothesis test, we need to find the sample-based estimate of the difference

$$\hat{\hat{d}} = \hat{\theta}_2 - \hat{\theta}_1$$

as well as its estimated standard error

$$\text{se}(\hat{\hat{d}}) = \text{se}(\hat{\theta}_2 - \hat{\theta}_1) = \sqrt{\text{var}(\hat{\theta}_2) + \text{var}(\hat{\theta}_1) - 2\,\text{cov}(\hat{\theta}_2, \hat{\theta}_1)}$$

From there, determining whether

$$t = \frac{\hat{\hat{d}}}{\text{se}(\hat{\hat{d}})}$$

is significantly different from 0 is tantamount to determining whether the parameter has changed significantly over time.

Generally speaking, unless the two point estimates are composed of data from a disjoint set of PSUs, then the estimated covariance term will be non-zero. Acquiring an estimate of the covariance term is somewhat of a nuisance, but we will demonstrate below a few ways to incorporate it implicitly—one using PROC SURVEYREG for comparisons of sample means and another using replicate weights generated from a replication variance estimation technique.

The remainder of this section is structured as follows. In Section 4.2, comparisons of means are discussed. In Section 4.3, we shift the focus to comparisons of totals. Finally, Section 4.3 sketches out a strategy involving replication variance estimation techniques, which is applicable to any population parameter of interest (including means and totals).

4.2 TESTING CHANGES IN MEANS

Perhaps the most common significance test desired is for changes in a mean over time. For example, within the context of the mathematics aptitude survey example, school officials may want to compare the mean year-end test score in 2017 to the year-end score in 2016. In other words, they want to determine whether the observed difference—hopefully an increase—is a statistically significant one.

A handy tool to use in this setting is PROC SURVEYREG. Why PROC SURVEYREG? Because fitting a simple linear regression model where the *y* variable is the outcome of interest and the lone *x* variable is an indicator variable coded 1 if the survey record is from the second year and 0 if the survey record is from the first year provides all of the information we need. Specifically, the test for whether the parameter associated with the 0/1 predictor variable is significantly different from 0 is equivalent to

$$t = \frac{\hat{\bar{y}}_2 - \hat{\bar{y}}_1}{\sqrt{\text{var}(\hat{\bar{y}}_2) + \text{var}(\hat{\bar{y}}_1) - 2\,\text{cov}(\hat{\bar{y}}_2, \hat{\bar{y}}_1)}}$$

where a subscript of 2 indicates the second survey time period and a subscript of 1 indicates the first survey time period.

The syntax below uses the stacked survey data set GRADES to test whether the mean grade at the end of the school year changed significantly in 2017 relative to 2016. Only the pertinent output is reported. We can interpret the intercept of 89.2 as the estimated mean year-end score in 2016—the average when YEAR=0—and the estimated parameter associated with YEAR to be the change in the estimated mean moving from YEAR=0 to YEAR=1, or from 2016 to 2017. Although the average score went up about 1.7 percentage points, the *t* statistic was not large enough in magnitude for statistical significance to be declared.

Note that the VADJUST=NONE option after the slash in the MODEL statement suppresses a correction factor of $(n - 1)/(n - p)$ all SURVEY procedures with linear modeling capabilities—PROC SURVEYREG, PROC SURVEYLOGISTIC, and PROC SURVEYPHREG—apply to entries of the estimated model parameter covariance matrix, which was originally proposed by Hidiroglou et al. (1980). When *n* is large relative to *p*, the adjustment factor is approximately 1 and therefore has little impact, but we have chosen to omit the correction factor in syntax examples appearing in this paper so that the algebra matches up exactly with what would be found from a two-sample *t* test conducted independently of PROC SURVEYREG.

Also note that additional factors can be controlled for by adding terms to the right of the equals sign in the MODEL statement of PROC SURVEYREG. For example, one could add demographic categories or other covariates such as whether or not the student received any kind of mathematics tutoring. The benefit in doing so is one of interpretation: rather than interpreting any observed difference in an overall sense, it would be interpreted as difference accounting for whichever covariates have been included in the model.

```
*** testing a change in the mean of a numeric variable;
proc surveyreg data=grades;
  stratum class;
  cluster homeroom;
  model grade2 = year / vadjust=none;
weight weight;
run;
```

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	89.2055944	1.38523620	64.40	<.0001
year	1.7136257	2.36277760	0.73	0.4822

Although it may seem unnatural at first, a linear regression model can also be exploited for significance tests of means of binary outcome variables (i.e., proportions). This is closely related to the concept of a risk difference discussed in Agresti (2013) and in Chapter 4 of Lewis (2016). For example, school officials may be interested in comparing the estimated proportion of students who pursue any kind of mathematics tutoring during the two school years. In the syntax below, an indicator variable TUTOR_Y is

created with a 1 to indicate the student received tutoring and 0 otherwise, which is treated as the outcome variable in a linear regression model with the 0/1 indicator variable YEAR as the single predictor variable. Based on the output, we can infer that an estimated 25.94% of students were tutored in 2016, whereas 25.94% + 18.97% = 44.91% were tutored in 2017. Although the increase appears sizeable in magnitude, the *t* statistic is not large enough to declare statistical significance.

```
*** testing a change in the mean of a 0/1 categorical variable;
data grades;
  set grades;
tutor_Y=(tutor='Y');
run;

proc surveyreg data=grades;
  stratum class;
  cluster homeroom;
  model tutor_Y = year / vadjust=none;
weight weight;
run;
```

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.25944056	0.08944410	2.90	0.0133
year	0.18972379	0.14584892	1.30	0.2177

4.3 TESTING CHANGES IN TOTALS

Another parameter for which significance tests are frequently desired is a finite population total. In the context of the mathematics aptitude survey, for example, school officials may be interested in comparing whether the total number of students seeking mathematics tutoring changed significantly between 2016 and 2017. Because the PSUs associated with the two school years are mutually exclusive, the covariance between the two year-specific point estimates is 0, and so we can extract everything we need to carry out this test by using the DOMAIN statement in PROC SURVEYMEANS—for more discussion on the topic of domain estimation, see Chapter 8 of Lewis (2016).

When PROC SURVEYMEANS sees a DOMAIN statement, it computes the statistics requested by the user for the full data set, and then repeats for subsets of the data defined by each distinct code of the DOMAIN statement variable(s). For example, the syntax below produces an overall table of statistics the entire stacked data set (output not shown), followed by the like for YEAR=0 and YEAR=1 which correspond to the 2016 and 2017 school years, respectively. The options SUM and VARSUM in the PROC SURVEYMEANS statement request that the estimated total and estimated variance for all domain categories be computed.

Based on the output, the estimated total number of students in the high school receiving mathematics tutoring in 2016 is 185.5, and 322.5 in 2017. Hence, the numerator of the test statistic is 322.5 – 185.5 = 137. The denominator is the square root of the sum of the two estimated variances of the point estimates, or SQRT(7074.25 + 14728) = 147.66. Because 137 / 147.66 is less than 2, prior to even referencing the appropriate student *t* distribution, we can anticipate that this will not be a significant difference.

```
*** testing for a change in a total;
proc surveymeans data=grades sum varsum;
  stratum class;
  cluster homeroom;
  class tutor;
  var tutor;
domain year;
weight weight;
run;
```

Domain Statistics in year

Year	Variable	Level	Sum	Std Dev	Var of Sum
0	tutor	N	529.500000	165.904340	27524
		Y	185.500000	84.108561	7074.250000
1	tutor	N	395.500000	136.024814	18503
		Y	322.500000	121.357942	14728

Again, it bears repeating that we can carry out this hypothesis test using output from PROC SURVEYMEANS because data from the two years (i.e., the two domains) do not emanate from one or more of the same PSUs—granted, we could have used PROC SURVEYFREQ to get the same figures (see Program 4.1 of Lewis (2016)). If this is not the case, then the standard error of the difference will generally contain a non-zero covariance, which PROC SURVEYMEANS (or PROC SURVEYFREQ) is not designed to estimate. Because the covariance is often positive, thinking back to the formulas presented in Section 4.1, it is advantageous to account for it, because it can serve to reduce the estimated standard error of the difference. One way to implicitly account for the covariance is via the general methodology shown in Section 4.4. In fact, the methodology works for differences in any kind of parameter, including means and totals.

4.4 TESTING CHANGES IN OTHER PARAMETERS

In this section, we introduce a general methodology one can use for testing changes in any kind of parameter (or any function of parameters) based on ideas of replication variance estimation strategies (Rust and Rao, 1996; Wolter, 2007). The premise of these replication techniques is to treat the data set as if it were the population and repeatedly sample from it in some systematic fashion. From each sample, or replicate, the estimate of interest is computed, and the variability of the estimate from the full data set is approximated by a simple function of the variability amongst the replicate-specific estimates. An appealing feature is that there is generally only one variance formula (per method), regardless the underlying quantity being estimated. The entire process can be efficiently implemented after appending a series of replicate weights to the analysis data set. As will be shown, PROC SURVEYMEANS can be used to create and append the replicate weights.

Without loss of generality, we will consider only one such strategy in this paper, the *jackknife* (see Chapter 4 of Wolter, 2007). There are actually several closely related forms of the jackknife used in practice, but the one we will demonstrate is the traditional method, what Valliant et al. (2008) refer to as

the “delete-one” version. Specifically, each PSU is deleted, in turn, and the remaining PSUs are weighted up to form a replicate-specific estimate.

The replicate weights are constructed as follows:

1. For units in the dropped PSU, all weights are set to 0.
2. For units in the same stratum as the dropped PSU, what SAS refers to as the *donor stratum*, the weights are inflated by a factor of $n_h/(n_h - 1)$, where n_h is the number of PSUs in the donor stratum.
3. For units outside the donor stratum, the replicate weight takes on the same value as the original weight. (If there is no stratification in the sample design, this step is skipped.)

After computing the full-sample point estimate, $\hat{\theta}$, and the point estimate using all replicate weights, $\hat{\theta}_r$ (r = 1, …, R), the jackknife variance estimate is

$$\text{var}_{JK}(\hat{\theta}) = \sum_{r=1}^{R} \frac{n_h - 1}{n_h} (\hat{\theta}_r - \hat{\theta})^2$$

The multiplicative factor $(n_h - 1)/n_h$ associated with each replicate is called the *jackknife coefficient*; it is identical for all replicates within the same stratum.

To motivate an example with respect to the high school mathematics aptitude survey, suppose the school officials are interested in estimating the ratio of the students’ year-end test score to the score they earned when taking the test at the beginning of the school year. Hopefully, the ratio is greater than 1. A ratio of, say, 1.05 indicates that students’ aptitudes, on average, increased 5% as measured by the test. While we could use the RATIO statement in PROC SURVEYMEANS to find this ratio and estimated standard error for a single school year, there is no direct way to find the estimated standard error for the difference of two ratios across two school years.

The syntax below demonstrates how to create and utilize jackknife replicate weights to tackle this analysis task. The purpose of the first PROC SURVEYMEANS run is not to conduct any particular analysis, only to create and append the jackknife replicate weights. The default Taylor series linearization method of variance estimation is changed to the jackknife by specifying the VARMETHOD=JK option in the PROC SURVEYMEANS statement. The version of the input data set GRADES with replicate weights appended is named GRADES_JK. Because the jackknife involves creating a replicate weight for each PSU dropped, and there are a total of 16 PSUs (8 from each year), a total of 16 replicate weights are appended. By default, they are named RepWt_1, RepWt_2,…., RepWt_16. The OUTJKCOEFS=COEFS_JK option in parentheses produces a data set with 16 observations, one for each jackknife coefficient needed to find each replicate weight estimate’s contribution to the overall variance estimate.

Similar in spirit the Program 9.14 of Lewis (2016), the %GETVAR macro begins by finding the full-sample estimate for the ratio difference, one based on the original analysis weight called WEIGHT, then progresses to find the same point estimates for all 16 replicate weights. The jackknife coefficients are merged into the JK_EST_ALL data set, and are multiplied by the squared deviations of each replicate-specific point estimate from the full-sample point estimate. A PROC SQL step towards the end of the macro definition outputs the original full-sample estimate and its estimated standard error, which is a function of the entries in the JK_EST_ALL data set. From the output, we find the estimated ratio of year-end average grade to that at the beginning of the school year was 1.0348 in 2017, which decreased from an estimated value of 1.0403 in 2016, but with a standard error estimated to be 0.012968, the difference is not a statistically significant one.

```
*** testing for a change in a Ratio via jackknife replicate weights;
* initial PROC SURVEYMEANS run to create replicate weights;
```

```
proc surveymeans data=grades varmethod=JK (outweights=grades_JK
                                           outJKcoefs=coefs_JK);
  stratum class;
  cluster homeroom;
  var grade2;
weight weight;
run;

* macro to estimate difference in ratio of grade2 to grade1 across years;
%macro getvar(replicates);
* housecleaning;
proc sql; drop table JK_est_all; quit;

* store full-sample estimates in macro variables;
proc sql;
  select sum(weight*grade2)/sum(weight*grade1) into :ratio1
    from grades_JK where year=1;
  select sum(weight*grade2)/sum(weight*grade1) into :ratio0
    from grades_JK where year=0;
quit;
data _NULL_;
ratio_diff_FS = &ratio1 - &ratio0;
call symput('ratio_diff_FS',ratio_diff_FS);
run;

* loop through all replicate weights and get the same estimate;
%do r=1 %to &replicates;
proc sql;
  create table JK_est_&r as
  select ratio1 - ratio0 as est from
    (select sum(RepWt_&r*grade2)/sum(RepWt_&r*grade1) as ratio1
      from grades_JK where year=1),
    (select sum(RepWt_&r*grade2)/sum(RepWt_&r*grade1) as ratio0
      from grades_JK where year=0);
quit;
proc datasets nolist;
  append data=JK_est_&r base=JK_est_all FORCE;
  delete JK_est_&r;
run;
quit;
%end;

* consolidate replicate-specific estimates and find squared deviations from
  full-sample estimate;
data JK_est_all;
  merge coefs_JK
        JK_est_all;
var=JKCoefficient*(est - &ratio_diff_FS)**2;
run;

* output point estimate and estimated standard error;
proc sql;
  select &ratio1 as Ratio1,&ratio0 as Ratio0,
         &ratio1 - &ratio0 as Difference,
         sqrt(sum(var)) as SE_JK
    from JK_est_all;
quit;
```

```
%mend;

%getvar(16);
```

Ratio1	Ratio0	Difference	SE_JK
1.040319	1.03475	0.005569	0.012968

This jackknife routine will work in other survey settings and for any parameters, although results from Kovar et al. (1988) suggest the jackknife should be avoided when estimating measures of variability for quantiles. As demonstrated in Section 9.7 of Lewis (2016), replication procedures are also applicable to linear model parameter estimates. Of course, alternative replication procedures such as the bootstrap and balanced repeated replication (Wolter, 2007) could be employed as well in a similar manner. While not necessarily algebraically equivalent, when carried out properly, all replication procedures provide valid, unbiased variance estimates.

Lastly, note that another appealing feature of the replicate weight approach is that the survey data set(s) can be subsetted for purposes of targeting particular sub-populations. For example, if school officials were interested in the ratio and/or ratio difference between the 2016 and 2017 school years only for students who received any kind of tutoring, the GRADES data set could be filtered in between the initial PROC SURVEYMEANS run and execution of the %GETVARS macro. In general, as discussed in detail in Chapter 8 of Lewis (2016), except in certain special circumstances, subsetting is not recommended when using the SURVEY family of procedures' default Taylor series linearization method for variance estimation

5. POOLING SURVEY DATA SETS TO INCREASE PRECISION

In this section, we sketch out two strategies that can be used when pooling two or more survey data sets for the purposes of formulating a single point estimate, a point estimate that can be interpreted as the midpoint of the *T* survey time periods. Often this is done to increase the sample size, and therefore precision, for an estimate of a relatively rare subpopulation.

Lee et al. (2007) sketch out two potential estimators. The first involves averaging the *T* point estimates by finding

$$\hat{\theta}_a = \frac{\sum_{t=1}^{T} \hat{\theta}_t}{T}$$

and then estimating the variance as

$$\text{var}(\hat{\theta}_a) = \frac{\sum_{t=1}^{T} \text{var}(\hat{\theta}_t)}{T^2}$$

The second estimator involves finding a weighted average based on the population sizes, N_t, associated with each time period. Specifically, the weighted average would be

$$\hat{\theta}_w = \frac{\sum_{t=1}^{T} N_t \hat{\theta}_t}{\sum_{t=1}^{T} N_t}$$

with variance

$$\text{var}(\hat{\theta}_w) = \left(\frac{N_t}{\sum_{t=1}^{T} N_t} \right)^2 \text{var}(\hat{\theta}_t)$$

If the N_t's, or the population sizes at the various survey time periods, do not vary much, there won't be much difference in the two estimators—for the special case where the N_t's are all equal, the two estimators are equivalent.

To motivate an example using the mathematics aptitude survey, rather than compare the proportions of students who receive tutoring across the two school years, maybe it would suffice for school officials to combine data from the two school years to report a single estimate, one that can be interpreted as the average rate of tutoring over the two-school-year period. The PROC SURVEYMEANS syntax below produces all necessary ingredients to formulate either estimator discussed in Lee et al. (2007). The SUMWGT option in the PROC statement outputs the sum of the weights for each year, which we can interpret as the N_t's, or at least their estimated values. Table 1 summarizes the point estimate and estimated standard error for the two estimators. Given that there is very little change in the school's population across the two years (N_1 = 715 and N_2 = 718), it is perhaps of little surprise that the two point estimates do not differ much. The same is true of the standard errors, but notice how the two combined estimators produce a smaller standard error than either of those estimated within a particular year.

```
*** pooling data to increase sample size;
proc surveymeans data=grades sumwgt mean;
  stratum class;
  cluster homeroom;
  class tutor;
  var tutor;
weight weight;
domain year;
run;
```

Domain Statistics in year

year	Variable	Level	Sum of Weights	Mean	Std Error of Mean
0	tutor	N	715.000000	0.740559	0.089444
		Y	715.000000	0.259441	0.089444
1	tutor	N	718.000000	0.550836	0.104788
		Y	718.000000	0.449164	0.104788

Estimator	Point Estimate	Estimated Standard Error
$\hat{\theta}_a = \frac{\sum_{t=1}^{T} \hat{\theta}_t}{T}$	(0.259441 + 0.449164) / 2 = 0.3543	SQRT[(1/4) x (0.089444^2 + 0.104788^2)] = 0.0689
$\hat{\theta}_w = \frac{\sum_{t=1}^{T} N_t \hat{\theta}_t}{\sum_{t=1}^{T} N_t}$	[(715)(0.259441) + (718)(0.449164)] / [715 + 718] = 0.3545	=SQRT[((715/(715+718))^2)*(0.089444^2) + ((718/(715+718))^2)*(0.104788^2)] = 0.0689

Table 1. Comparison of Two Pooled Survey Data Set Estimators for the Proportion of Students Receiving Mathematics Tutoring Using the Hypothetical Mathematics Aptitude Survey Data.

Technically, if one or more of the same PSUs appear across any of the *T* years, then there are (likely positive) covariance terms that factor into either estimated variance formula shown above. In contrast to the difference examples covered in Section 4, a positive covariance does not serve to decrease the overall variance. Instead, it would serve to increase it, as both of the estimator's variance estimates are a function of the sum of the *T* year-specific variances. However, a replication variance estimation strategy comparable to the one demonstrated in Section 4.4 could be employed to incorporate the covariance terms implicitly. Essentially, the idea would be to determine the given estimators independently for each replicate weight and consolidate the results as we did earlier with the %GETVAR macro. For brevity, we will not walk through an example of that in this paper.

6. CONCLUSION

This paper outlined several estimation strategies analysts might find useful when combining two or more survey data sets. Two general analysis tasks were distinguished: (1) pooling survey data to conduct significance tests of point estimate changes over time; and (2) pooling survey data to arrive at a single, more precise point estimate. The latter is typically pursued when a single survey year data set does not contain enough data to make reliable inferences about a particular subpopulation of interest. Pooling two or more data sets together can increase the sample size and, thus, precision, but it changes the interpretation to mean the point estimate at or around the midpoint in time of the repeated survey administrations. This may be acceptable, or even preferred, depending on the context of the analysis and how much the underlying target population changes across the two or more time periods.

REFERENCES

Agresti, A. (2013). *Categorical Data Analysis, Third Edition*. New York, NY: Wiley

Hidiroglou, M., Fuller, W., and Hickman, R. (1980). *Super Carp*. Ames, IA: Statistical Laboratory, Iowa State University.

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

Kovar, J., Rao, J.N.K., and Wu, C. (1988). "Bootstrap and Other Methods to Measure Errors in Survey Estimates," *Canadian Journal of Statistics*, 16, 25–45.

Lee, S., Davis, W., Nguyen, H., McNeel, T., Brick, J.M., and Flores-Cervantes, I. (2007). "Examining Trends and Averages Using Combined Cross-Sectional Survey Data from Multiple Years," CHIS Methodology Paper.

Lewis, T. (2016). *Complex Survey Data Analysis with SAS®*. Boca Raton, FL: CRC Press.

Rust, K., and Rao, J.N.K. (1996). "Replication Methods for Analyzing Complex Survey Data," *Statistical Methods in Medical Research: Special Issue on the Analysis of Complex Surveys,* **5**, pp. 283 – 310.

Valliant, R., Brick, J.M., and Dever, J. (2008). "Weight Adjustments for the Grouped Jackknife Variance Estimator," *Journal of Official Statistics*, **24**, pp. 469 – 488.

Wolter, K. (2007). *Introduction to Variance Estimation*. *Second Edition*. New York, NY: Springer.

CONTACT INFORMATION

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

Taylor Lewis
Department of Statistics
George Mason University
tlewis18@gmu.edu
http://mason.gmu.edu/~tlewis18/