Paper 5181-2016

The Use of Statistical Sampling in Auditing Health-Care Insurance Claim Payments

Taylor Lewis, Julie Johnson, and Christine Muha, U.S. Office of Personnel Management[1]

ABSTRACT

This paper is a primer on the practice of designing, selecting, and making inferences on a statistical sample, where the goal is to estimate the magnitude of error in a book value total. Although the concepts and syntax are presented through the lens of an audit of health-care insurance claim payments, they generalize to other contexts. After presenting the fundamental measures of uncertainty that are associated with sample-based estimates, we outline a few methods to estimate the sample size necessary to achieve a targeted precision threshold. The benefits of stratification are also explained. Finally, we compare several viable estimators to quantify a book value discrepancy, making note of the scenarios where one might be preferred over the others.

BACKGROUND

One of the many tasks undertaken by the Office of the Inspector General (OIG) of the U.S. Office of Personnel Management (OPM) is to conduct periodic reviews of questionable health-care insurance claim payments made from within the Federal Employee Health Benefits (FEHB) program overseen by OPM. For instance, Report Number 1A-99-00-14-046 (OPM, 2015) describes an audit of questionable claims identified as part of the global coordination of benefits for a large nationwide health-care insurance corporation. The purpose of the audit was to determine whether, and to what extent, the organization complied with contract provisions relative to coordination of benefits with Medicare. Historically, the scope for many of these audits has been restricted to high-dollar claim amounts (i.e., those greater than or equal to $2,500), with each and every questionable claim being reviewed. That is, high-dollar claim amounts are censused, enabling the book value discrepancy to be measured with certainty. Recently, efforts have been made to employ statistical sampling on the relatively more voluminous low-dollar claims (i.e., those less than $2,500) to estimate its share of the total book value discrepancy. The purpose of this paper is to outline some of the methods, considerations, and SAS® tools the team utilizes for designing and conducting these audits involving statistical sampling.

The paper begins with an overview of the fundamental concepts and terminology associated with sampling to make inferences about some larger population. After laying that groundwork, we demonstrate several methods to estimate a book value discrepancy in a simple random sampling context, and comment on a few strategies for sample size determination. Examples illuminate how auxiliary information can be exploited to improve sampling efficiency, either during the design stage, the estimation stage, or both.

KEY CONCEPTS AND TERMINOLOGY FOR SAMPLE DESIGN AND ESTIMATION

The first task in any sampling effort is to define the *target population* about which inferences are desired. The target population often carries an ambitious, all-encompassing label, such as “the general U.S. population” or “all car accidents occurring in Arlington County, Virginia during calendar year 2015.” The next step is to construct a list, or *sampling frame*, from which a random sample of *sampling units* can be drawn. The totality of entities covered by this list is called the *survey population*, which does not always coincide perfectly with the target population, as illustrated by Figure 1. The area of the target population falling outside the survey population area is of most concern. That region represents *undercoverage*, or a portion of the target population that has no chance of being selected into the sample. Sometimes it is possible to supplement one sampling frame with another to capture this group, or conduct weighting adjustment techniques such as *poststratification* or *raking* (Lewis, 2012) to compensate for it during the

[1] The opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect those of the U.S. Office of Personnel Management.

estimation stage. Undercoverage does not necessarily mean that the sampling effort is a nonstarter, but any known issues should be should acknowledged as a limitation.

Notice how there is also an area in Figure 1 delineating a portion of the survey population falling outside the bounds of the target population. This represents extraneous, ineligible sampling units on the sampling frame that may be selected as part of the sample. This is a much easier problem to deal with than undercoverage. We will walk through a simple example of how to do so later in the paper.

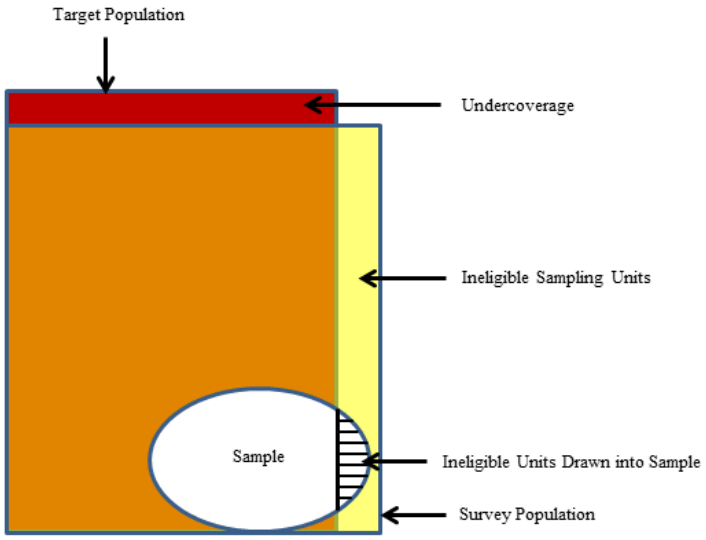


Figure 1. Visualization of a Sample Relative to the Target Population and Survey Population.

Another fundamental task is to identify the *population parameters* to be estimated from the sample. These are often symbolized by Greek letters, or Roman letters without a "hat." Sample-based estimates are typically distinguished by topping them with a "hat." For example, the population mean of some continuous variable *y* might be denoted μ or $\bar{y}$, whereas a sample-based estimate of the quantity might be denoted $\hat{\mu}$ or $\hat{\bar{y}}$. You want to be as specific as possible when identifying population parameters of interest. For example, in the health-care insurance claims setting, we know the total book value of claims paid, which we can symbolize *X*, but we do know not how much *should* have been paid, *Y*. Perhaps the primary objective of the audit is to produce an estimate of this quantity, $\hat{Y}$. Given these two quantities, we can obtain an estimate of the total book value discrepancy, $\hat{D} = X - \hat{Y}$. The crux of this paper is to demonstrate alternative methods for estimating $\hat{D}$. All methods are unbiased, valid approaches; however, depending on how the data pairs (x_i,y_i) for the $i = 1, 2, \ldots, N$ population units are structured, some are more efficient than others.

Given an interest in a generic population parameter, θ, the sample-based estimate, $\hat{\theta}$, is typically referred to as a *point estimate*. Generally speaking, we would not expect the point estimate from a sample of size *n* from a population of size *N* to equal the population parameter precisely (unless perhaps $n = N$, meaning we conducted a census, or sampled the entire the population). Rather, we would anticipate that a different sample would produce a somewhat different point estimate. This is the notion of *sampling error*. Interestingly, however, we can estimate the magnitude of sampling error using data from only one sample.

The basic building block of any measure of sampling error is the estimated *variance* of the point estimate, $\text{var}(\hat{\theta})$. This can be interpreted as an estimate of the average squared deviation of point estimates from the true population parameter if we were to repeat the given sampling procedure many times. The variance formula to be employed is a function of three factors:

The population parameter of interest – for example, the variance of an estimated mean has a different formula than an estimated total.

The sample design – for example, the sample size, population size, and whether or not the sampling frame was stratified (more discussion forthcoming).

The *estimator*, or mechanism used to formulate a point estimate – as we have alluded to, there are multiple viable estimators for a population total.

As an aside, there is another class of variance estimators collectively referred to as *replication techniques* (Rust and Rao, 1996) based on the notion of estimating sampling error by treating the sample as if it were the population and repeatedly sampling from it in some systematic fashion. See Lewis (2015) for a more in-depth discussion of these approaches along with SAS syntax examples.

Numerous measures of uncertainty can be derived from the variance. For one, the square root of the variance is called the *standard error* of the point estimate. In formulas, $\text{se}(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta})}$. We can use the standard error to compute the *coefficient of variation* (CV) of a point estimate, defined as $\text{CV}(\hat{\theta}) = \frac{\text{se}(\hat{\theta})}{\hat{\theta}}$. This is synonymously referred to as the *relative standard error* (RSE), as it relates the magnitude of sampling error to the magnitude of the point estimate itself. For example, a CV of 0.25 means that the standard error of a point estimate is 25% the size of the point estimate itself. The most appealing feature of the CV is that it is unit-less. It can be used to make precision comparisons between two or more unique population parameters (e.g., a mean versus a total) or even two different estimators of the same population parameter.

Another popular way to employ a point estimate's standard error is to construct a (1 – *α*)100% *confidence interval* (CI) as $\hat{\theta} \pm t_{df,1-\alpha/2}\text{se}(\hat{\theta})$. Technically, we are free to choose *α*, the significance level, yet it is quite common for analysts to assign *α* = 0.05, which corresponds to a 95% CI. Indeed, this is the default in most SAS procedures. If two values, *a* and *b*, represent the lower and upper endpoints of a 95% CI, then we can say we are 95% confident that the true population parameter falls somewhere between *a* and *b*. A smaller *α* leads to a larger value of $t_{df,1-\alpha/2}$, the $(1 - \alpha/2)^{\text{th}}$ percentile of a student *t* distribution with *df* degrees of freedom and, hence, a wider CI. The degrees of freedom are a function of the sample design. In a simple random sampling setting, $df = n - 1$, but other rules may apply for other designs. For example, if the sample design involves stratification with *H* distinct strata (more discussion forthcoming), then $df = n - H$. SAS calculates *df* and $t_{df,1-\alpha/2}$ for us; we merely need to concern ourselves with properly alerting it of all applicable sample design features.

Finally, yet another derivative measure of uncertainty worth mentioning is the *margin of error* of a point estimate, which is one-half the width of the CI associated with it. In other words, the margin of error is the term to the right of the ± symbol in the CI formula: $\text{moe}(\hat{\theta}) = t_{df,1-\alpha/2}\text{se}(\hat{\theta})$.

AN EXAMPLE DATA SET

To facilitate exposition of the theory and execution of sampling strategies illustrated in this paper to estimate a book value discrepancy, let us introduce an example data set. The data set CLAIMS is a population of 7,650 questionable health-care claims that have been exhaustively audited. It includes the following four variables:

- ID_claim – numeric identifier of a unique claim, ranging from 1 to 7,650
- X – dollar amount of claim actually paid
- Y – dollar amount of claim that *should* have been paid
- D – difference in what was paid and what *should* have been paid

Despite the fact that we have information on the entire target population, thereby negating the need to conduct an audit sample, let us suppose that the objective is to estimate $D = X - Y$, the total book value

discrepancy, by reviewing 100 claims. We will assume the book value total of the entire target population is fixed and known as X = \$1,812,595, but that we must estimate Y from the sampled claims.

The syntax in Program 1 illustrates how we can use PROC SURVEYSELECT to draw a simple random sample of n = 100 claims from the population of N = 7650 claims. All options appear in the PROC statement. For example, the sampling frame is identified in the DATA= option, METHOD=SRS requests simple random sampling (without replacement), and the sample size is specified by the N=100 option. The output data set SAMPLE_SRS named in the OUT= option consists of 100 records from the CLAIMS data set chosen at random. (Alternatively, we could specify the option OUTALL, in which case the full input data set is output with an indicator variable SELECTED flagging the sampled records.) Providing a number in the SEED= option ensures the exact same sample of claims will be produced if we need to submit the same code at a later point in time. Lastly, the STATS option forces the probability of selection (n/N = 100/7650 = 0.013) and the sampling weight, the inverse of the probability of selection (N/n = 76.5) to be output in the columns SELECTIONPROBABILITY and SAMPLINGWEIGHT, respectively. The sampling weight can be conceptualized as the number of population units represented by each sampled case. For example, a weight of 5 indicates the sampling unit represents itself and 4 others just like it in the population. As we will see in forthcoming syntax examples, the sampling weights are used to amplify results from the sample to the population.

Program 1. Selecting a Simple Random Sample (SRS) without Replacement of 100 Claims.

```
proc surveyselect data=claims n=100 seed=23488829 method=SRS stats
                  out=sample_SRS;
run;
```

Figure 2 below provides a glimpse of the first few records of the data set SAMPLE_SRS. We see that the first claim sampled was the 72nd, the second claim sampled was the 350th, and so on.

VIEWTABLE: Work.Sample_srs

	ID_claim	x	y	d	Probability of Selection	Sampling Weight
1	72	131.82	131.82	0	0.0130718954	76.5
2	350	54.5	0	54.5	0.0130718954	76.5
3	383	5.58	5.58	0	0.0130718954	76.5
4	413	9.05	0	9.05	0.0130718954	76.5
5	485	76.55	0	76.55	0.0130718954	76.5

Figure 2. Partial View of Data Set SAMPLE_SRS.

Using this (simulated) sample of audited claims, in the next section we demonstrate four distinct estimation strategies for the total book value discrepancy, $\hat{D} = X - \hat{Y}$. Each strategy is unbiased, but estimates of precision can vary widely depending on how the data are structured. After walking through syntax examples, we briefly summarize with a visually-aided discussion contrasting the implied models underlying the strategies, in an effort to illuminate the particular data structures where a given strategy will be produce the most precise point estimate.

ESTIMATORS OF A BOOK VALUE DISCREPANCY

1. THE MEAN-PER-UNIT ESTIMATOR

The first and most straightforward strategy we can consider is the *mean-per-unit* (MPU) estimator. The idea is to first estimate the mean amount that should have been paid from the sample of claims, and then multiply it by the total number of claims in the population. In formulas, this is to say we find $\hat{Y}_{MPU} = N\hat{\bar{y}}$. Given this, we can estimate the total book value discrepancy as $\hat{D}_1 = X - \hat{Y}_{MPU}$.

The estimated variance of this quantity is $\text{var}(X - \hat{Y}_{MPU}) = \text{var}(\hat{Y}_{MPU}) = N^2 \text{var}(\hat{\bar{y}})$ where

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

Equation 1 represents the variance of a sample mean given a simple random sample without replacement of size *n* from a population of size *N*. The middle term is the estimated variance of the y_i's, and is often symbolized s^2. Note that the estimated standard deviation of the y_i's, often symbolized *s*, is just the square root of the variance. The rightmost term represents a *finite population correction* (FPC), which converges to 0 as *n* approaches *N*. Essentially, this term credits us with a precision increase whenever we sample a non-negligible portion of the population. In the event *n* = *N*, the FPC is 0, which means the variance is 0. This makes sense. As we have all information, there is no sampling error. Lastly, note that the FPC disappears from Equation 1 whenever sampling is done with replacement, and sometimes it is purposefully ignored if *n*/*N* is only trivially greater than 0. Purposefully ignoring it leads to a slight overestimation of variability, but it simplifies computations.

Program 2 demonstrates how to use PROC SURVEYMEANS to find $\hat{Y}_{MPU}$ and its standard error. We point SURVEYMEANS to the sampling weight on the input sample data set, and specify the keyword SUM to indicate we want to estimate the total(s) for the variable(s) specified in the VAR statement. The TOTAL=7650 option tells PROC SURVEYMEANS the value of *N*, which is needed to incorporate the FPC into estimated measures of variability.

Program 2. The Mean-per-Unit (MPU) Estimator of a Total.

```
proc surveymeans data=sample_SRS total=7650 sum;
  var y;
weight SamplingWeight;
run;
```

```
                    The SURVEYMEANS Procedure

                          Data Summary

              Number of Observations           100
              Sum of Weights                  7650


                        Statistics

Variable              Sum         Std Dev
-----------------------------------------
y                 1240630          233777
-----------------------------------------
```

Using the output, we can reason that the estimated book value discrepancy is \$1,812,595 - \$1,240,630 = \$571,965, with a standard error of \$233,777. Note that the standard error of a total in PROC SURVEYMEANS output is labeled "Std Dev," which is somewhat of a misnomer. This will likely be modified in a future version of SAS.

Sometimes the sampling frame contains ineligible, or out-of-scope, population units unable to be removed beforehand. If we find such units in our sample, we must account for them during the estimation process. As an example, suppose we determined during the audit that the claims associated with ID #'s 383, 2002,

6108, and 7121 were all ineligible (e.g., misclassified, from the wrong time period). Because we found 4/100 = 4% of our sample claims to be ineligible, it is reasonable to expect that approximately 4% of claims in the full sampling frame are ineligible. Following terminology of Lohr (2009), the eligible portion of the population can be referred to as a *domain*. Although it seems natural to target this domain of interest by simply subsetting the SAMPLE_SRS data set for only the 96 eligible claims, this is risky. The proper method is to provide the full data set to PROC SURVEYMEANS, but create an indicator of eligibility and specify it in the DOMAIN statement. Subsetting will not affect the point estimate, but measures of variability may not be correct. The simplest explanation is that PROC SURVEYMEANS assumes n = 96, when in truth n = 100, but see Lewis (2013a) for a more in-depth discussion.

Program 3 below demonstrates how to account for the ineligibility. PROC SURVEYMEANS produces an overall point estimate, and an estimate for all subsets of the data identified by unique codes of the DOMAIN statement variable(s). Without working out the algebra of the revised book value discrepancy estimate, we can immediately infer from the output that the estimate would drop by $91,571 after excluding the four ineligible cases. The standard error of the estimate drops a bit as well, down from $233,777 to $230,130.

Program 3. Accounting for Ineligibility of a Subset of Sampled Cases when Estimating a Population Total Using the MPU Estimator.

```
data sample_SRS;
  set sample_SRS;
eligible=(ID_claim not in(383 2002 6108 7121));
run;

proc surveymeans data=sample_SRS total=7650 sum;
   var y;
weight SamplingWeight;
domain eligible;
run;
```

```
                    The SURVEYMEANS Procedure

                Domain Statistics in eligible

eligible    Variable             Sum         Std Dev
-----------------------------------------------------
       0    y                  91571           61560
       1    y                1149059          230130
-----------------------------------------------------
```

This exercise of dealing with ineligibility was done only for purposes of instruction. The remaining examples in this section will treat the full 100 cases in the SAMPLE_SRS data set as eligible.

Although we are operating at present under the assumption of a fixed sample size of 100 claims, in many circumstances we seek to determine the sample size needed to meet a maximum tolerable value for a particular measure of uncertainty. In general, this can be accomplished by proceeding through the following steps:

1. Define the population parameter you are interested in estimating.
2. Choose the measure of uncertainty you want to target.
3. Choose the maximum acceptable value for the chosen measure of uncertainty.
4. Write out the expression for the measure of uncertainty.
5. Solve for n by inserting values for all other inputs and rearrange such that n sits on one side of the

equation.

The first four steps are the easy part. The tricky part is determining the inputs for the fifth step and carrying out the algebra to solve for *n*. Both can necessitate some creative thinking.

Suppose the goal is to estimate *D* via the MPU estimator for *Y*, and that we want to do our best to ensure the standard error of the estimate is no greater than \$100,000, which would correspond to a 95% CI of approximately ± \$200,000. In this case, our measure of uncertainty can be expressed as $\text{se}(X - \hat{Y}_{MPU}) = \sqrt{N^2 \text{ var}(\hat{\bar{y}})}$. Substituting in what we know about Equation 1 and the fact that *N* = 7650, the task is to solve the following inequality for *n*: $100000 < \sqrt{7650^2 \left(\frac{s^2}{n}\right)\left(1 - \frac{n}{7650}\right)}$.

We still have one extra unknown to contend with: s^2, the variance of the y_i's. Two plausible strategies to get a handle on this quantity: (1) conduct a small-scale pilot study of, say, 30 cases (you can always lump these cases back in with the remaining cases from the audit sample); or (2) exploit findings from a prior audit. As one example of the second strategy—based on the authors' actual experience—suppose we knew from a previous audit that approximately 40% of claims are found to be paid in error such that the amount that *should* have been paid is \$0, whereas the other 60% were paid correctly. In this situation, we can simulate sample data to get a gauge on the variance of the y_i's. Program 4 illustrates example syntax to do so. Without going over each and every detail, the idea behind the approach is to simulate 100 hypothetical samples and generate pseudo-y_i's governed by this assumed relationship. From there, we find 100 simulated values of s^2. We use the average of these simulated values as our input for sample size determination purposes. From the final PROC MEANS output (not shown), this average value comes out to \$92,939.71.

Program 4. An Example of Simulating the Observed Data for Purposes of Estimating the Variance Input to a Sample Size Determination Problem.

```
* replicate 100 copies of the sampling frame;
proc surveyselect data=claims out=claims2 reps=100 sampsize=7650
                  seed=458047 method=SRS;
run;

* set approx. 40% of values to 0;
data claims2;
  set claims2;
y_sim=(ranuni(322211)>.4)*x;
run;

* get the variance for each replicate;
proc means data=claims2 noprint nway;
  class replicate;
  var y_sim;
output out=variances var=var_y_sim;
run;

* find the mean simulated variance;
proc means data=variances mean;
  var var_y_sim;
run;
```

Given an input for the s^2 term, it is just a matter of algebraic manipulation to solve for *n*. Depending on the complexity of the expression, this can be a tedious task. A handy work-around is to create a SAS data set calculating the measure of certainty for all possible sample sizes. The syntax to do so is demonstrated in Program 5. For values of *n* ranging from 2 to 7,650, an estimated standard error is calculated given all other known or assumed inputs. We can scroll down through the data set SES to find

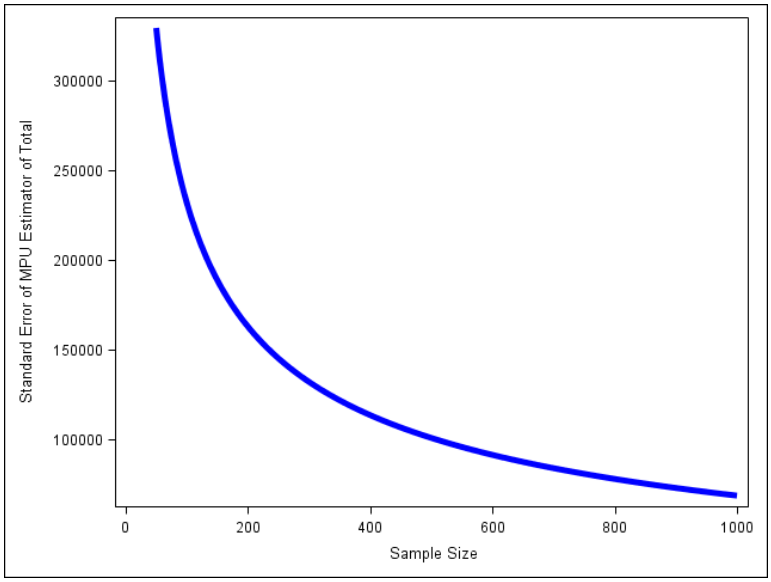
the value of the sample size column where SE_TOTAL first drops below 0.05. In the present case, that value is n = 510.

Another utility of iterating over all possible sample sizes is that we can plot the given measure of uncertainty as a function of the sample size. This is the carried out by the PROC SGPLOT syntax in Program 5. The plot shows how returns are relatively larger in the beginning but gradually diminish: the SE reduction is much greater increasing the sample size from 100 to 200 than it is increasing it from 500 to 600. Note that this tactic can be extended to two or more curves. For instance, it could prove insightful to overlay two curves corresponding to two assumed values of s^2.

Program 5. Determining the Necessary Sample Size by Computing and Plotting a Measure of Uncertainty for All Possible Sample Sizes.

```
* syntax to iterate through all sample sizes;
data SEs;
pop=7650;     * <-- population size;
s2=93262.71; * <-- assumed variance of outcome variable;
do n=2 to pop;
   SE_total=sqrt(pop**2*(s2/n)*(1 - n/pop));
output;
end;
run;

* plot the curve;
proc sgplot data=SEs;
  where 50 le n le 1000;
  series X=n Y=SE_total / lineattrs=(color=blue thickness=5);
xaxis label='Sample Size';
yaxis label='Standard Error of MPU Estimator of Total';
run;
quit;
```



Another great resource for sample size determination is Valliant et al. (2013), who extend the basic concepts discussed in this paper to multi-criteria optimization problems using PROC NLP in SAS/OR® and the (free) Microsoft Excel® Solver add-in.

2. THE DIFFERENCE ESTIMATOR

A second estimation strategy is to formulate a *difference estimator* (Guy et al., 2002). The idea is to create a variate $d_i = x_i - y_i$ for the i = 1, 2, …, n sampled claims, and then find the MPU estimator for the D. In formulas, this is to say that we find $\hat{\bar{d}} = \frac{\sum_{i=1}^{n} d_i}{n}$, and use it to estimate the total book value discrepancy as $\hat{D}_2 = N\hat{\bar{d}}$, a point estimate with variance $\text{var}(N\hat{\bar{d}}) = N^2 \text{var}(\hat{\bar{d}})$.

The mechanics of the difference estimator are very similar to those of the MPU estimator. As Program 6 demonstrates, the only difference is that we substitute into our PROC SURVEYMEANS code the variable D, the difference between variables X and Y, which is already stored on the SAMPLE_SRS data set. From the output, we find that the estimated book value discrepancy is $490,521 with a standard error of $131,663. Hence, this approach produces a markedly more precise estimator than the MPU estimator of Y.

Program 6. The Difference Estimator of a Total.

```
proc surveymeans data=sample_SRS total=7650 sum;
  var d;
weight SamplingWeight;
run;
```

```
                    The SURVEYMEANS Procedure

                          Data Summary

                Number of Observations           100
                Sum of Weights                  7650


                        Statistics

Variable              Sum          Std Dev
-----------------------------------------
d                  490521           131663
-----------------------------------------
```

3. THE RATIO ESTIMATOR

A third avenue for estimating the total book value discrepancy is to formulate a *ratio estimator* (Cochran, 1977). The notion is to first estimate from the sampled claims the ratio of dollars paid correctly to the ratio of dollars paid. In statistical notation, this is to say we determine $\hat{R} = \frac{\hat{Y}}{\hat{X}}$, and use it to estimate the total amount that should have been paid as $\hat{Y}_{ratio} = X\hat{R}$. From here, the total book value discrepancy can be estimated by $\hat{D}_3 = X - \hat{Y}_{ratio}$. Note that since $\text{var}(\hat{D}_3) = \text{var}(\hat{Y}_{ratio})$, the variance of this point estimate reduces to $X^2 \text{var}(\hat{R})$.

Program 7 demonstrates the SAS syntax to estimate $\hat{Y}_{ratio}$ and its standard error, which is also the standard error of the total book value discrepancy estimate. The trick is to create a new variable that is each value of y_i multiplied by X, and use that as the numerator of a ratio estimated by PROC

SURVEYMEANS. From the output, we see that $\hat{Y}_{ratio} = \$1{,}298{,}997$, which means that the total discrepancy estimate is \$1,812,595 - \$1,298,997 = \$513,598 with standard error \$131,423.

Program 7. The Ratio Estimator of a Total.

```
data sample_SRS;
  set sample_SRS;
numerator=1812595*y;
run;

proc surveymeans data=sample_SRS total=7650 ratio;
  ratio numerator / x;
weight SamplingWeight;
run;
```

```
                    The SURVEYMEANS Procedure

                         Ratio Analysis

Numerator Denominator        Ratio         Std Err
-----------------------------------------------------
numerator x                1298997          131423
-----------------------------------------------------
```

4. THE REGRESSION ESTIMATOR

Lohr (2009, p. 138) points out that ratio estimation works well if the relationship between the x_i and y_i pairs abide by an approximately straight-line pattern intersecting the y-axis at the origin (i.e., when x_i is zero, y_i is zero). If the pairs follow an approximately straight-line pattern, but not one that necessarily intersects the y-axis at the origin, we can instead formulate a *regression estimator*. The first step is to estimate a simple linear regression model from the sample data as $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i$, and then use it to estimate the mean of the y_i's in the population as $\hat{\bar{y}}_{reg} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, where $\bar{x}$ is the known mean claim amount paid in the population. From here, the total amount that should have been paid is estimated as $\hat{Y}_{reg} = N\hat{\bar{y}}_{reg}$, and so the estimated total book value discrepancy estimate is $\hat{D}_4 = X - \hat{Y}_{reg}$.

At first glance, this approach looks similar to the MPU estimator of Y, but the variance is a function of squared deviations from the regression line, which could be dramatically smaller than the variance of the y_i's if the there is a strong correlation between the x_i's and y_i's. Specifically, the variance of the point estimate reduces to $\text{var}(\hat{D}_4) = \text{var}(N\hat{\bar{y}}_{reg}) = N^2 \text{var}(\hat{\bar{y}}_{reg}) = N^2 \frac{s_\varepsilon^2}{n}\left(1 - \frac{n}{N}\right)$, where s_ε^2 is the mean squared error (MSE) of the regression model.

Program 8 illustrates syntax to obtain all ingredients to formulate a regression estimator for the total amount that should have been paid. First, note that since X = \$1,812,595 and N = 7,650, $\bar{x} = 236.94$. From the PROC SURVEYREG output, we find that the estimated intercept of the regression model is -21.516452, the estimated slope is 0.811732, and the MSE of the regression is $(163.12)^2$. Therefore, we can find $\hat{\bar{y}}_{reg} = -21.516453 + (0.811732)236.94 = 170.8158$, which means that the regression estimator for the estimated total amount that should have been paid is (7650)(170.8158) = \$1,306,740 with standard error $\sqrt{7650^2 \frac{(163.12)^2}{100}\left(1 - \frac{100}{7650}\right)} = \$123{,}969$. And so the estimated total book value

discrepancy is estimated as $\hat{D}_4 = X - \hat{Y}_{reg}$ = \$1,812,595 - \$1,306,740 = \$505,855, with the same standard error, \$123,969.

Program 8. The Regression Estimator of a Total.

```
proc surveyreg data=sample_SRS;
  model y = x;
weight SamplingWeight;
run;
```

The SURVEYREG Procedure

Fit Statistics

R-Square	0.7216
Root MSE	163.12
Denominator DF	99

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-21.516452	21.3159689	-1.01	0.3152
x	0.811732	0.1406674	5.77	<.0001

SUMMARY OF ESTIMATORS

Table 1 summarizes the four distinct point estimates and standard errors of the total book value discrepancy formulated from the same simple random sample of n = 100 claims. The point estimate for the MPU estimator of Y is the largest, but it also has the largest standard error. The three point estimates associated with the other three estimators, those that make use of auxiliary information—namely, the health-care insurance claim amount originally paid—are clustered closer to one another around the \$500,000 mark, and all have a much smaller standard error.

Estimator	Point Estimate	Standard Error
D_1 – MPU Y	\$571,965	\$233,777
D_2 – Difference Estimator	\$490,521	\$131,663
D_3 – Ratio Estimator	\$513,598	\$131,423
D_4 – Regression Estimator	\$505,855	\$123,969

Table 2. Summary of Point Estimates and Estimated Standard Errors from the Four Estimators of the Total Book Value Discrepancy from the Simple Random Sample of 100 Claims.

Figure 3 is provided to help visually contrast the implicit model underpinning each of these four estimators. Each of the four panels contain a scatterplot of the x_i's versus the y_i's in the data set SAMPLE_SRS overlaid with a red line intended to represent the underlying model exploited. For example, we can conceptualize the MPU estimator as a zero-slope model $y_i = \hat{\bar{y}} + \varepsilon_i$. On the other hand, the difference estimator uses a model $y_i = x_i + \varepsilon_i$, which we can interpret as a straight line intersecting the origin at a 45^0 angle. The ratio estimator can also be interpreted as a straight line intersecting the origin, but not necessarily at a 45^0 angle. Rather, the model is $y_i = \hat{R}x_i + \varepsilon_i$. Finally, the regression estimator can be interpreted as a straight line that does not necessarily pass through the origin, as it employs a model of the form $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i$. Generally, the most precise estimator will be the one with the smallest variance of the ε_i's, which is to say the one with the smallest average squared vertical

deviation between the observed y_i's and the red line. All things considered, it is no surprise that the MPU estimator was least precise, because it clearly exhibits deviations of the largest magnitude.

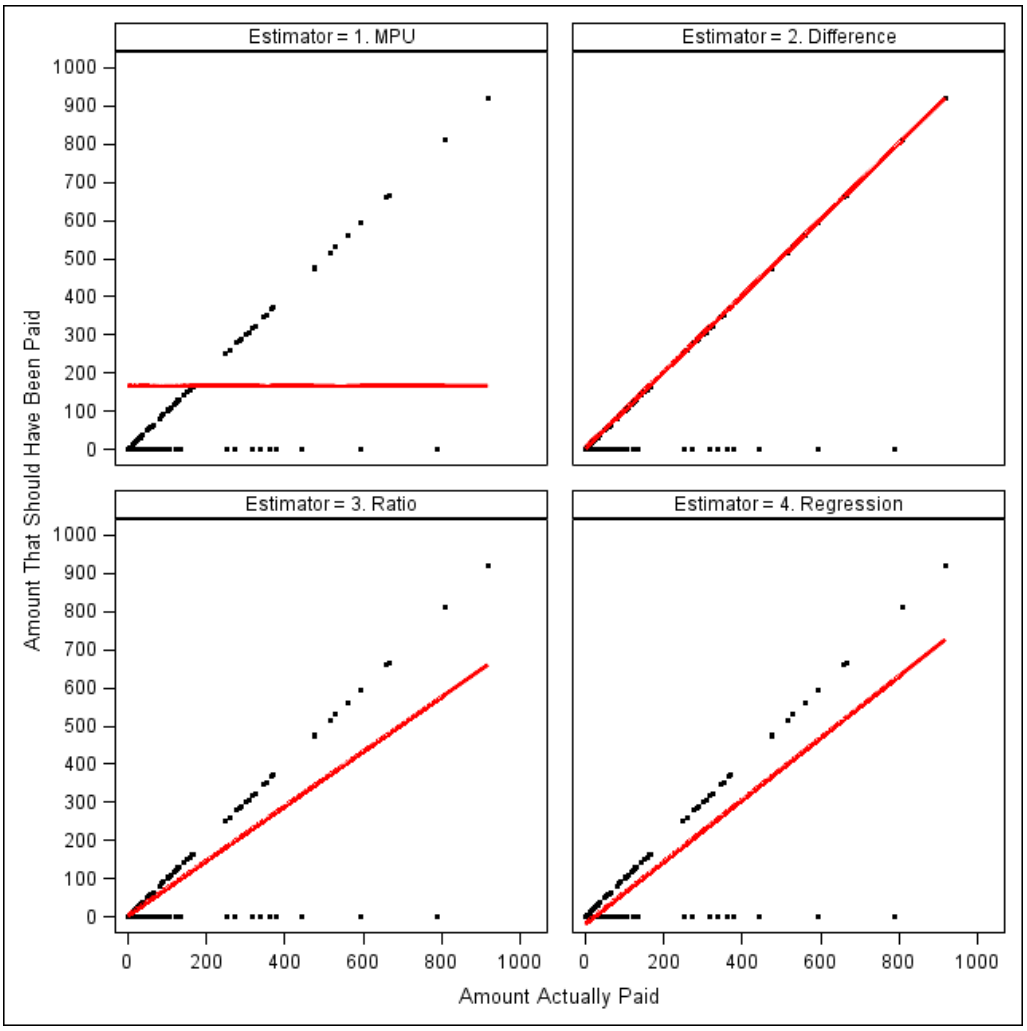


Figure 3. Scatterplot of Amount Paid versus Amount That Should Have Been Paid in the Sample of Claims Overlaid with a Reference Line Representing the Implicit Model Underlying the Estimator.

The main takeaway message is that you should plot the outcome variable(s) in the sample as a function of any fully observed auxiliary variable(s) available on the sampling frame to see what type(s) of relationships exists. Creating panels of scatterplots in the spirit of Figure 3 can provide insight into which types of estimators would be most appropriate and/or most precise. Although it probably goes without saying, there is nothing wrong with using different auxiliary variables and/or different estimators in making inferences about different population parameters from a single sample.

ALTERNATIVE SAMPLE DESIGNS

STRATIFICATION

Regardless of the estimator one ultimately uses to estimate a total book value discrepancy, a prudent technique one can use at the sample design stage is *stratification* (Cochran, 1977). Stratification involves partitioning the sampling frame into H mutually exclusive and exhaustive *strata* (singular: *stratum*), and then sampling independently within each. A few of the reasons stratification is used in practice:

- *Ensure representation of less prevalent subgroups in the population*. If there is a rare subgroup in the population that can be identified on the sampling frame, it can be sequestered into its own stratum to

provide greater control over the number of units sampled. Sometimes the subgroup is so small that it makes most sense just to census those units.

- *Administer multiple methods of data collection*. Within the purview of the health-care insurance claims audit, we can consider claims greater than $2,500, those subject to the exhaustive review and perhaps a somewhat different auditing procedure, as its own stratum, whereas claims less than $2,500 constitute a separate stratum (or set of strata).
- *Increase precision of point estimates*. When strata are constructed homogeneously with respect to the key outcome variable(s), substantial precision gains can be achieved.

Because sampling is performed independently within each stratum, we are effectively able to eliminate the between-stratum component of variability. To see how, consider the estimated variance of the overall sample mean under this sample design:

$$\text{var}(\hat{\bar{y}}_{st}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \text{var}(\hat{\bar{y}}_h) \quad (2)$$

where N_h is the stratum-specific population size, n_h is the stratum-specific sample size, and s_h^2 is the stratum-specific element variance. We can conceptualize this as the variance of a weighted sum of stratum-specific sample means, where weights are determined by the proportion of the population covered by the given stratum, or $\frac{N_h}{N}$, where $\sum_{h=1}^{H}\frac{N_h}{N} = 1$.

Ideally, one would use the outcome variable itself as the stratification factor; however, this is generally unknown. (If it were known, there would likely be no need to conduct a sample!) So the idea is to exploit an auxiliary variable from the sampling frame, one related to the outcome variable, in hopes that it helps "explain away" some of the inherent variability.

With respect to audit sampling, it is very common to define strata as dollar amount ranges. There is no single "best" method for creating these groupings. Cochran (1968) cites empirical evidence that most precision gains taper after about 5 strata. It is also important to note that any precision gains to be achieved are dependent on the population parameter being estimated. For example, Kish (1965) cautions that gains tend to be much less dramatic for means of population attributes (i.e., traits or characteristics) than for continuously measured variables.

Returning to our goal of estimating the book value discrepancy of questionable health-care insurance claims, Program 9 illustrates the precision gains we could achieve—maintaining the same marginal sample size of n = 100—by sampling n_h = 25 claims from each of H = 4 strata defined by the following dollar amount ranges: (1) $0 - $100; (2) $100 - $250; (3) $250 - $1000; and (4) $1,000 - $2,500.

The first part of the program consists of syntax to select the stratified sample. A stratum indicator variable is created on the sampling frame, and then the sampling frame is sorted by this indicator, which is placed in the STRATA statement of PROC SURVEYSELECT. Note that, because we must specify stratum-specific sample sizes, the SAMPSIZE= option points to a supplemental data set containing the stratum indicator code and a key column _NSIZE_. Also note that, because the strata do not consist of equal numbers of claims, there are variable sampling rates across strata. These variables rates of representation are compensated for during the estimation process by making use of the stratum-specific values of the SAMPLINGWEIGHT variable produced by PROC SURVEYSELECT.

The second part of the program consists of syntax to estimate the total amount that should have been paid from the stratified sample via the MPU estimator of Y. Recall that the FPC is a stratum-specific quantity, and so a single value passed to the TOTAL= option will generally not suffice. Instead, we can create and specify a supplemental data set with the stratum indicator code and a key column _TOTAL_ housing the N_h's. The only other difference in the PROC SURVEYMEANS code relative to Program 2 is that the stratum indicator is specified in the STRATA statement.

From the output, we find that the estimated discrepancy is $1,812,595 - $1,232,562 = $580,033 with standard error $120,247, which is much less than the standard error of $233,777 determined from the MPU estimator in the unstratified sample of the same size.

Program 9. Designing, Selecting, and Analyzing a Stratified Sample to Formulate an MPU Estimator or a Total.

```
* define four strata on the sampling frame;
data claims;
  set claims;
if 0 < x <= 100 then stratum=1;
  else if 100  < x <= 250  then stratum=2;
  else if 250  < x <= 1000 then stratum=3;
  else if 1000 < x <= 2500 then stratum=4;
run;

* create a supplemental data set with stratum-specific sample sizes;
data sampsizes;
  input stratum _NSIZE_;
datalines;
1 25
2 25
3 25
4 25
;
run;

* sort the sampling frame by stratification variable;
proc sort data=claims;
  by stratum;
run;

* select a stratified sample of size 25 x 4 = 100;
proc surveyselect data=claims n=sampsizes seed=5399255 method=SRS
                  out=sample_STR;
strata stratum;
run;

* create a supplemental data set with stratum population counts;
proc freq data=claims;
  table stratum /
        out=counts_strata (keep=stratum count
                           rename=(COUNT=_TOTAL_));
run;

* estimate the total claim amount that should have been paid;
proc surveymeans data=sample_STR total=counts_strata sum;
  var y;
weight SamplingWeight;
strata stratum;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Strata	4
Number of Observations	100
Sum of Weights	7650

Statistics

Variable	Sum	Std Dev
y	1232562	120247

The same prescriptions and techniques for targeting measures of uncertainty discussed previously are applicable for particular strata within a stratified sample designs. That is, you can approach the sample size determination problem at the stratum level and let the overall sample size be a function of stratum-specific necessary sample sizes. Another perspective is target an overall point estimate measure of uncertainty by, say, setting up an inequality with respect to the quantities specified in Equation 2. One potential quandary with this approach is that there may be an infinite number of solutions.

A popular alternative perspective to take in sample size determination problems for stratified designs is to begin with a fixed *n*, perhaps one dictated by budgetary constraints, and solve for the allocation amongst strata that minimizes some measure of uncertainty in the overall point estimate. One pertinent application is *Neyman allocation* (Lohr, 2009), which minimizes the variance of the overall sample mean (i.e., as specified in Equation 2) according to the following allocation formula:

$$n_h = n \times \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h} \tag{3}$$

This is pertinent for the present book value discrepancy problem because minimizing the variance of the overall sample mean is tantamount to minimizing the variance of the MPU estimator, considering $\text{var}(\hat{Y}_{MPU}) = N^2 \text{var}(\hat{\bar{y}}_{st})$. Note that Equation 3 is specified in terms of the outcome variable's standard deviations (σ_h's) with respect to the strata (sub)populations. These are generally unknown, but we can plug in estimates, or s_h's, using one of the strategies discussed earlier in the paper.

Program 10 shows how we can create a supplemental data set and use the ALLOC= option after the slash in the STRATA statement of PROC SURVEYSELECT to perform Neyman allocation to divvy up our fixed sample size of 100 claims amongst the four pre-defined strata. Not shown here, suppose that the simulation approach from Program 4 was fleshed out to get the stratum-specific variance inputs. Be advised that PROC SURVEYSELECT is looking for variances stored in a column named _VAR_, not standard deviations as specified in Equation 3. Specifying the NOSAMPLE option makes the output data set consist of stratum-specific sample sizes according to the given allocation procedure—for a discussion of other allocation procedures, see Lewis (2013b). It is good practice to carry out this preliminary step prior to actually selecting the sample and moving forward with data collection, just so you can verify that the resulting sample sizes are reasonable. As an example of something that can go awry, the allocation procedure may return a sample size of $n_h = 1$, which will cause problems for variance estimation purposes, because the default formulas used in PROC SURVEYMEANS require $n_h \geq 2$.

Figure 4 provides a view of the output data set SAMPSIZES_NEYMAN. Despite the lowest dollar range stratum having the greatest number of claims in the population, it would be sampled at the lowest rate.

This is because the variance of the outcome variable is much smaller in this stratum relative to the three higher-dollar range strata.

Program 10. Illustration of Neyman Allocation of a Fixed Sample Size to a Pre-Defined Set of Strata for Minimizing the Variance of the Overall Sample Mean.

```
data stratum_vars;
  input stratum _VAR_;
datalines;
1    925
2   5500
3  77859
4 590584
;
run;

proc surveyselect data=claims out=sampsizes_neyman sampsize=100;
  strata stratum / alloc=neyman var=stratum_vars nosample;
run;
```

VIEWTABLE: Work.Sampsizes_neyman

	stratum	Total Number of Sampling Units	Variance	Allocation Proportion	Sample Size	Actual Proportion of Total Sample Size
1	1	3946	925	0.116600191	12	0.12
2	2	1536	5500	0.1106735587	11	0.11
3	3	1779	77859	0.4822826884	48	0.48
4	4	389	590584	0.2904435619	29	0.29

Figure 4. View of Neyman Allocation Results from PROC SURVEYSELECT.

PROBABILITY PROPORTIONAL TO SIZE SAMPLING

Lastly, yet another method for employing a continuous auxiliary variable believed to be highly correlated with the outcome variable is to use it as the measure of size in a *probability proportional to size* (PPS) sample design (Lohr, 2009). PPS sampling is sometimes referred to synonymously as *dollar-unit sampling* in auditing contexts, so-named because each dollar in however the population is defined has an equal chance of being part of the sample.

The notion behind a PPS design is to ascribe to each sampling unit on the sampling frame the following probability of selection:

$$\Pr(\text{unit } i \text{ sampled}) = n \times \frac{x_i}{\sum_{i=1}^{N} x_i} \qquad (4)$$

where x_i is the measure of size for the i^{th} sampling unit. The net effect is that units with larger measures of size are more likely to be sampled, and *vice versa*. As with other designs, we can employ the SAMPLINGWEIGHT variable computed by PROC SURVEYSELECT during the estimation stage to compensate for these variable probabilities of selection. Though we will not delve into the theory, the potential benefit of PPS sampling is that, if the x_i's and y_i's are strongly correlated with one another, we can estimate Y with much more precision.

Program 11 shows syntax to select a PPS sample (without replacement) of n = 100 claims using the METHOD=PPS option in PROC SURVEYSELECT. The measure of size is X, the claim amount originally paid, which is specified in the SIZE statement. From there, PROC SURVEYMEANS syntax to estimate the total amount that should have been paid mirrors what we have seen previously. From the output, we can gather that the estimated total book value discrepancy is estimated as \$1,812,595 - \$1,248,878 = \$563,717 with standard error \$83,599. Hence, at least with the data at hand, this turns out to be the most precise estimator yet.

Program 11. Probability Proportional to Size (PPS) Sampling and Estimation of a Total.

```
proc surveyselect data=claims out=sample_PPS method=PPS
                  seed=24999232 stats n=100;
  size x;
run;

proc surveymeans data=sample_PPS total=7650 sum;
  var y;
weight SamplingWeight;
run;
```

```
                    The SURVEYMEANS Procedure

                          Data Summary

                Number of Observations           100
                Sum of Weights             9829.73098


                          Statistics

Variable              Sum         Std Dev
-----------------------------------------
y                 1248878           83599
-----------------------------------------
```

Be advised when using METHOD=PPS that PROC SURVEYSELECT will stop executing if any unit's share of the total measure of size is greater than 1/*n*. Referring back to Equation 4, this causes the probability of selection to equal or exceed 1, which is impossible. If faced with this situation, a sensible work-around would be to place all units where $\frac{x_i}{\sum_{i=1}^{N} x_i} \geq \frac{1}{n}$ in their own "certainty" stratum, where they can be censused. Indeed, this is the default strategy in the auditing software EZ-Quant (http://www.dcaa.mil/ez_quant_applications.html).

CONCLUSION

We began this paper by introducing the terminology and fundamental quantities for measuring the uncertainty in a point estimate of a population parameter derived from a statistical sample. We then illustrated SAS syntax to conduct several commonly utilized strategies for estimating a book value discrepancy. These techniques are useful, if not imperative, for auditing large accounts, those where a full review would be prohibitively expensive. Regardless of the specific technique(s) chosen, the main takeaway message is that one can markedly increase the precision of point estimates of population totals by utilizing an auxiliary variable correlated with it. The auxiliary variable can be employed either during the sample design stage, as we saw with stratification and PPS sampling, or during the estimation stage, as we saw with the difference, ratio, and regression estimators.

REFERENCES

Cochran, W. (1968). "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, **24**, pp. 295 – 313.

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

Guy, D., Carmichael, D., and Whittington, R. (2002). *Audit Sampling: An Introduction. Fifth Edition*. New York, NY: Wiley.

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

Lewis, T. (2012). "Weighting Adjustment Methods for Nonresponse in Surveys." *Invited paper presented at the Western Users of SAS Software (WUSS) Conference*. Long Beach, CA, September 5 – 7. Available online at: www.wuss.org/proceedings12/162.pdf.

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

Lewis, T. (2013b). "PROC SURVEYSELECT as a Tool for Drawing Random Samples." *Paper presented at the Southeast SAS Users Group (SESUG) Conference*. St. Pete Beach, FL, October 20 – 23. Available online at: http://analytics.ncsu.edu/sesug/2013/SD-01.pdf.

Lewis, T. (2015). "Replication Techniques for Variance Approximation." *Invited paper presented at the SAS Global Forum*. Dallas, TX, April 26 – 29. Available online at: http://support.sas.com/resources/papers/proceedings15/2601-2015.pdf.

Lohr, S. (2009). *Sampling: Design and Analysis. Second Edition*. Boston, MA: Brooks/Cole.

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.

U.S. Office of Personnel Management, Office of the Inspector General. (2015). "Audit of Global Coordination of Benefits for BlueCross and BlueShield Plans," Report No. 1A-99-00-14-046. Available online at: https://www.opm.gov/our-inspector-general/reports/2015/audit-of-global-coordination-of-benefits-for-bluecross-and-blueshield-plans-1a-99-00-14-046.pdf.

Valliant, R., Dever, J., and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York, NY: Springer.

CONTACT INFORMATION

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

Taylor Lewis
Mathematical Statistician
Survey Analysis Group
U.S. Office of Personnel Management
Taylor.Lewis@opm.gov