

An Introductory Overview of the Features of Complex Survey Data

Taylor Lewis, Joint Program in Survey Methodology (JPSM), University of Maryland

ABSTRACT

A “complex” survey data set is one characterized by any combination of the following four features: stratification, clustering, unequal weights, or finite population correction factors. In this paper, context is provided as to why these features may appear in data sets produced from surveys, some of the formulaic modifications they introduce are highlighted, and the syntax needed to properly account for them is outlined. Specifically, we explain why you should utilize the SURVEY family of SAS/STAT® procedures, such as PROC SURVEYMEANS or PROC SURVEYREG, to analyze data of this type.

1. INTRODUCTION

In the era of “Big Data,” the variety of statistics that can be generated is ostensibly limitless. Given the copious and ever-expanding types of data being collected, there are many questions that can be answered from analyzing a data set already in existence or perhaps even one updated in real time. For instance, a credit card issuer seeking to determine the total amount of charges made by its customers on gas during a particular year may have this information readily retrievable from one or more databases. If so, a straightforward query can produce the answer. On the other hand, determining the average amount the typical U.S. household spends on gas is a much more complicated estimation problem. Collecting data from all households in the United States would obviously be exceedingly costly, if not an outright logistical impossibility. One could probably make some progress pooling together the comprehensive set of credit card issuers’ databases and trying to group into distinct households via the primary account holder’s address, but not all households own and use a credit card, and this would necessarily omit any non-credit-card payment such as one made by cash or check. A survey of the general U.S. population is needed to acquire this kind of information with acceptable precision. One such survey is the Consumer Expenditure Survey, sponsored by the Bureau of Labor Statistics, which reported in September 2013 that the average household spent approximately \$2,500 on gas during calendar year 2012 (<http://www.bls.gov/news.release/cesan.nr0.htm>).

As authoritative or “official” as these statistics seem, it is important to bear in mind they are *estimates*. Note that the term *estimate* can sometimes be confused with very similar term *estimator*, but the two terms have different meanings. An estimate is the value computed from a sample, whereas an estimator is the method or technique used to produce the estimate. If the entire survey process were conducted anew, we would expect estimates to differ somewhat for a variety of reasons. Using formal statistical theory and a single sample data set in hand, however, there are established ways to calculate an unbiased estimate of the sampling error, which can be reported alongside the estimate or used to form of a confidence interval. A distinctive aspect of *complex* survey data, the features of which will be detailed in Section 4, all too often overlooked by applied researchers is that the techniques and formulas for estimating sampling error one learns in general statistics courses or from general statistics textbooks frequently do not carry over intact. The reason is that complex surveys employ alternative sample designs, either for purposes of statistical efficiency or out of necessity to control data collection costs. The implied data-generating mechanism in general statistics courses is simple random sampling with replacement of the ultimate units for which data is measured. In applied survey research, that particular data-generating mechanism is the exception rather than the rule.

This paper is structured as follows. Section 2 establishes some of the terminology pertaining to applied survey research. Section 3 previews the SAS/STAT procedures that have been developed to facilitate complex survey data analysis. These are all prefixed with the word SURVEY (e.g., PROC SURVEYMEANS is the companion procedure to PROC MEANS). And Section 4 introduces the four features that may be present in a survey data set justifying the qualifier “complex”: (1) finite population corrections; (2) stratification; (3) clustering; and (4) unequal weights.

2. DEFINITIONS AND TERMINOLOGY OF SAMPLE SURVEYS

Groves et al. (2009, p. 2) define a survey as a “systematic method for gathering information from (a sample of) entities for the purposes of constructing quantitative descriptors of the attributes of the larger population for which the entities are members.” They use the term “entities” to stress the fact that, although the word “survey” often has the connotation of an opinion poll or a battery of questions directed at humans, this is not always the case. Other example entities are farms, businesses, or even events. Parenthetically including the phrase “a sample of” serves to remind us that not all surveys involve sampling. A *census* is the term describing a survey that aims to collect data on or enumerate an entire population.

One of the first stages of any survey sampling effort is defining the *target population*, the “larger population” alluded to in the Groves et al. definition about which inferences are desired. The target population often carries an ambitious, all-encompassing label, such as “the general U.S. population.” 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 mesh perfectly with the target population. For example, there is no population registry in the United States as there is in many European countries to serve as a sampling frame. There is an important distinction to be made between the sampling units and the *population elements*, or the ultimate analytic units for which measurements are taken and inferences drawn. The two are not always one and the same. Sometimes the sampling frame consists of clusters of the population elements. Considering the goal of making inferences on the general U.S. population, even if a population registry existed, it might be oriented by household or address instead by individual. This would present a cluster sampling situation, which is permissible but introduces changes to the more familiar formulas used for estimation. We will discuss cluster sampling more in Section 4.4 with the help of a simple example.

A sampling frame’s makeup is often influenced by the survey’s *mode*, or method of data collection (e.g., in-person interview or self-administered paper questionnaire by mail). For example, a popular method for administering surveys by telephone is *random-digit dialing* (RDD), in which the sampling frame consists of a list of landline telephone numbers. A survey opting for this mode may still consider the target population “the general U.S. population,” but the survey population is actually the subset of U.S. households with a landline telephone.

Figure 1 below illustrates how the target population and survey population may not always coincide. The area within the target population that does not fall within the survey population area is of most concern. That represents *undercoverage*, meaning population elements that have no chance of being selected into the sample. Continuing with the RDD example, households without a landline telephone fall within this domain. Sometimes it is possible to supplement one sampling frame with another to capture this group (e.g., incorporating a sampling frame consisting of cellular telephone numbers), but that can introduce duplicated sampling units (i.e., households with landline *and* cellular numbers, therefore present in both frames), which can be a nuisance to deal with in its own right (Lohr and Rao, 2000). Another remedy often pursued is to conduct weighting adjustment techniques such as poststratification (Holt and Smith, 1979) or raking (Izrael et al., 2000).

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, or ineligible, sampling units on the sampling frame that may be selected as part of the sample. In terms of the example RDD survey of U.S. households, a few such possibilities are inoperable, unassigned, or business telephone numbers. These are represented by the area to the right of the vertical line in the oval labeled “Sample.” An appreciable rate of ineligibility can cause inefficiencies in the sense that these units must be “screened” out where identified, but that situation is generally easier to handle than undercoverage.

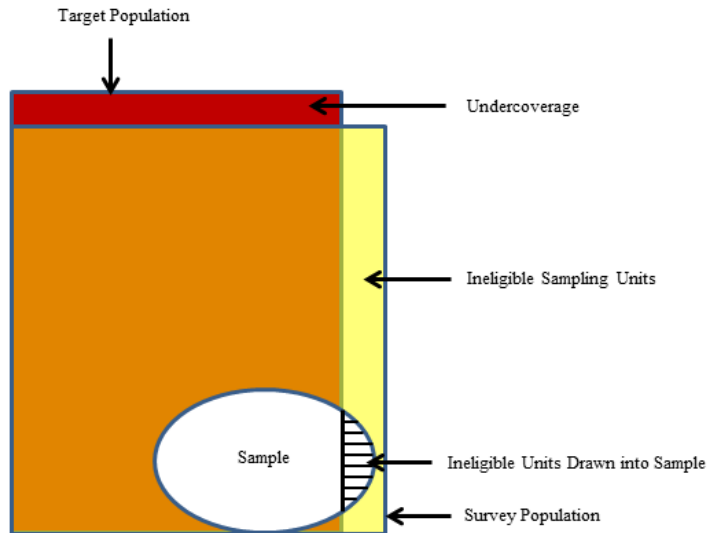


Figure 1: Visualization of a Sample Relative to the Target Population and Survey Population

It is not unusual for those unfamiliar with survey sampling theory to be skeptical at how valid statistical inferences can be made using a moderate sample of size n from a comparatively large population of size N . This is a testament to the *central limit theorem*, which states that the distribution of a sequence of means computed from independent, random samples of size n is always normally distributed so long as n is “sufficiently large.” Despite the vague descriptor “sufficiently large” fostering a certain amount of debate—some might say 25, others 30, and still others 50—the pervading virtue of the theorem is that it is true regardless of the underlying variable’s distribution. In other words, you do not need normally distributed data for this theorem to hold. This assertion is best illustrated by simulation.

Figure 2 shows the distribution of three variables from a fabricated finite population of $N = 100,000$. The first variable is normally distributed with a population mean of $\bar{y}_1 = 5$. The second is right-skewed with a mean of $\bar{y}_2 = 2$, whereas the third variable has a bimodal distribution with a mean of $\bar{y}_3 = 3.75$. Figure 3 immediately following displays the result of a simulation that involved drawing 5,000 samples of size $n = 15$, $n = 30$, and $n = 100$ from this population and computing the sample mean for each of y_1 , y_2 , and y_3 . That is, the figure is comprised of histograms summarizing the distribution of the three variables’ sample means with respect to each of the three sample sizes. As in Figure 2, the row factor distinguishes the three variables, while the column factor (moving left to right) distinguishes the increasing sample sizes. There are a few key takeaways from examining Figure 3:

- All sample mean distributions closely resemble a normal distribution, which has been superimposed on the histograms. Again, this is true regardless of the underlying distribution.
- The average, or *expected* value of the 5,000 sample mean values is the population mean, which is to say the sample mean is an unbiased estimate for the mean of the entire population.
- The distributions get “skinnier” as the sample size increases. This reflects increased precision, or less deviation amongst the means from one sample to the next.

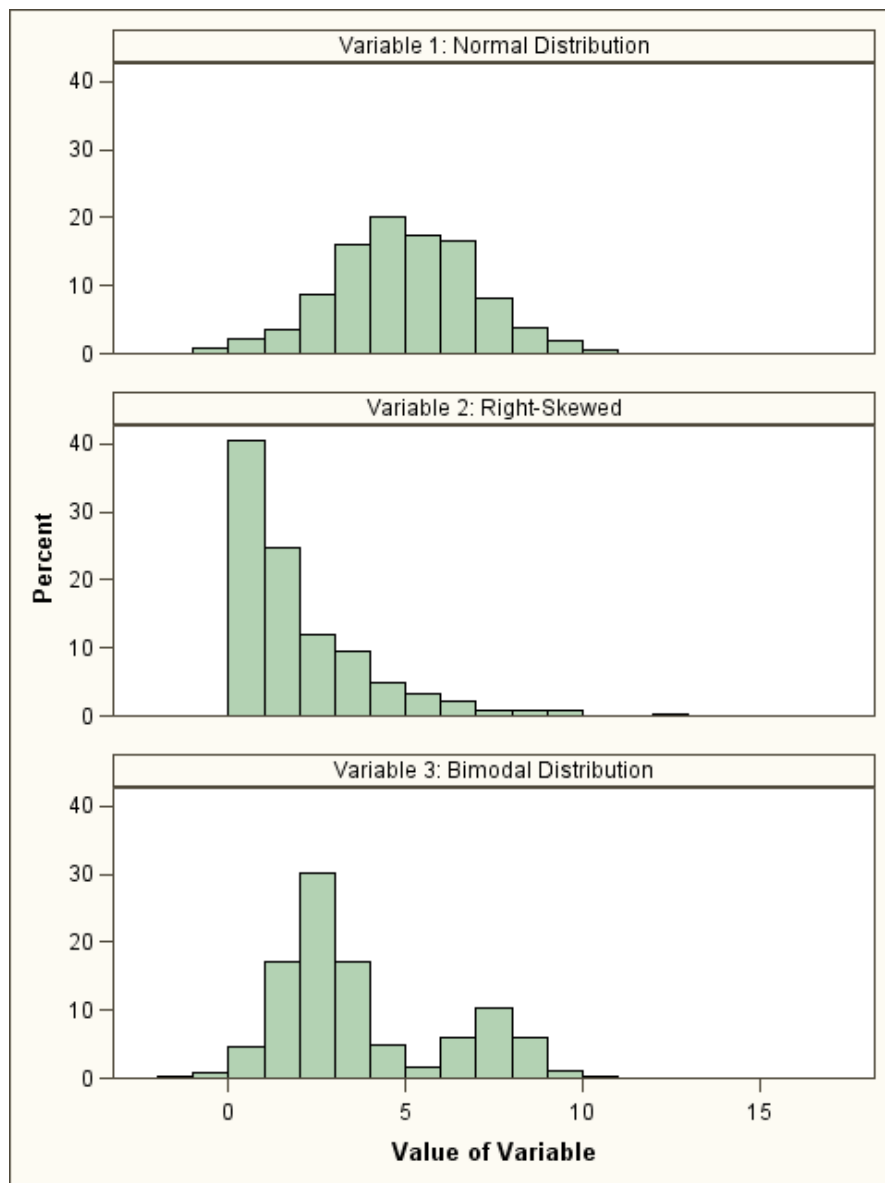


Figure 2: Distribution of Three Variables from a Finite Population of $N = 100,000$

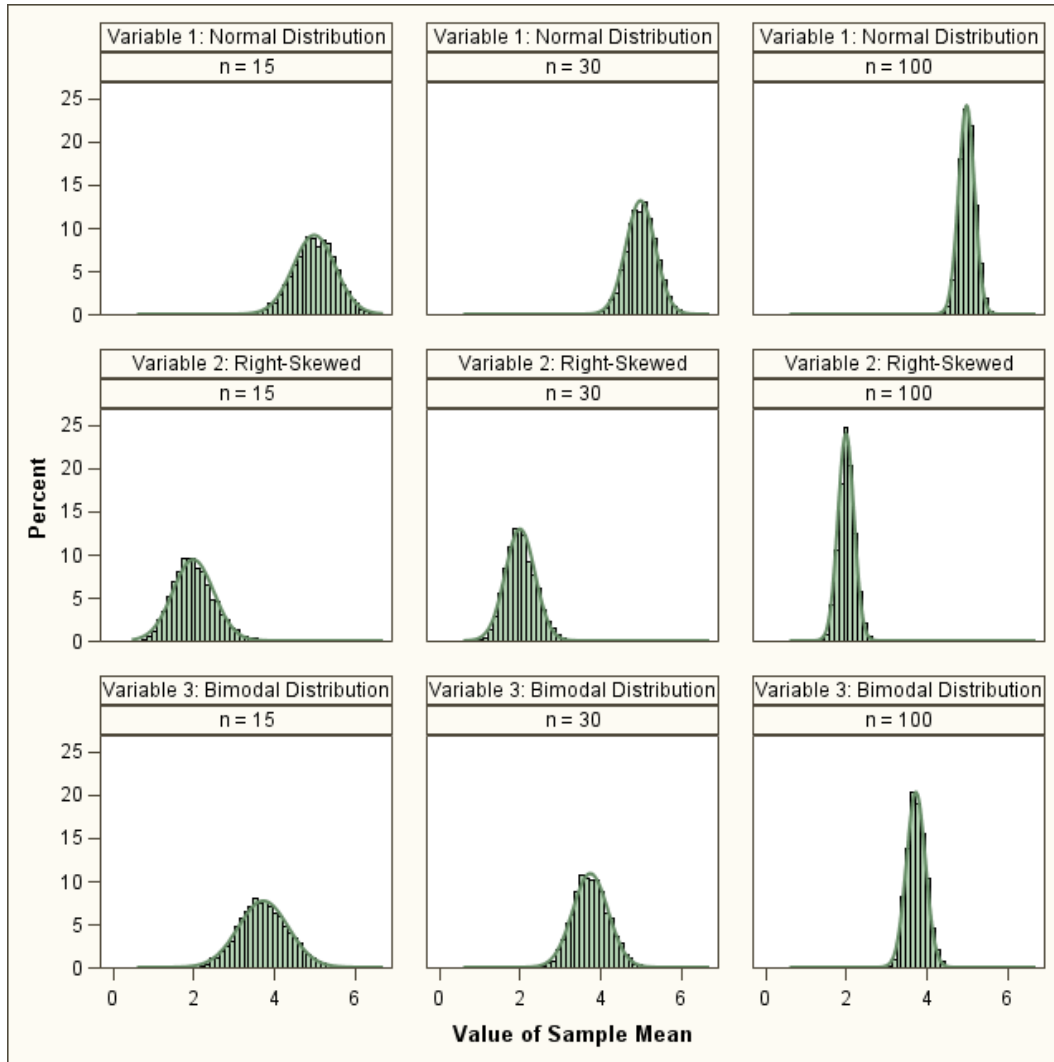


Figure 3: Sample Mean Distributions of 5,000 Samples of Size 15, 30, and 100 Drawn from the Finite Population in Figure 2

Knowing the sampling distribution of statistics such as the mean is what justifies our ability to make inferences about the larger population (e.g., to form confidence intervals, conduct significance tests, and calculate p values). In practice, we do not have a set of 5,000 independent samples, but usually a single sample to work with. From this sample a statistic is produced, which we typically top with a hat to distinguish it from the true population parameter we are estimating. In the general θ notation, we denote $\hat{\theta}$ as the sample-based estimate or *point estimate* of θ . For example, $\hat{\bar{y}}$ refers to the point estimate of \bar{y} from a particular sample. We acknowledge, however, that this estimate likely does not match exactly the true population value. A fundamental quantification of this anticipated deviation is the *variance*, which can be expressed as $\text{Var}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$. The variance represents the average squared deviation of the sample-based estimates from the true population value over all possible samples. This quantity is rarely known, except perhaps in simulated settings such as the one just discussed. But drawing from formal statistical theory, there are established computational methods to formulate an unbiased estimate of it with the single sample at hand. The sample-based estimate is referred to as the *estimated variance* or *approximated variance*, and denoted $\text{var}(\hat{\theta})$ (with a lower-case “v”).

Numerous related measures of uncertainty can be derived from the estimated variance. Since variability in squared units can be difficult to interpret, a common transformation made is to take the square root of this quantity, which returns the *standard error* of the statistic denoted $se(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta})}$. Along with the estimate itself, the standard error can be used to form a confidence interval $\hat{\theta} \pm t_{df, \alpha/2} se(\hat{\theta})$, where $t_{df, \alpha/2}$ represents the $100(1 - \alpha/2)^{\text{th}}$ percentile of a t distribution with df degrees of freedom. Another popular measure is the relative standard error of the estimate, or the *coefficient of variation*, which is defined as $CV(\hat{\theta}) = \frac{se(\hat{\theta})}{\hat{\theta}}$. This has an appealing interpretation. A value of 0.1, for example, indicates the standard error of the statistic represents 10% of the magnitude of the statistic itself. A target CV can be used as a sample design criterion before the survey is administered or as a threshold for whether a survey estimate is deemed precise enough to be published. Unlike the variance and standard error, the CV is unit-less, thereby permitting precision comparisons regardless of scale (e.g., dollars versus percentages) or type of estimate (e.g., means versus totals).

It should be emphasized that the lines of reasoning above only apply if the sample is selected randomly. An impressive-sounding sample size of, say, 10,000 means little if it were drawn using a non-random process such as convenience sampling, quota sampling, or judgment sampling (see p. 5 of Lohr, 2010). The essential requirement is that every sampling unit on the sampling frame has a known, non-zero chance of being selected as part of the sample. The selection probabilities need not be equal, but they should be fixed and non-zero.

3. OVERVIEW OF SAS/STAT PROCEDURES AVAILABLE TO ANALYZE SURVEY DATA

Table 1 summarizes the SURVEY procedures currently available in SAS® as of Version 9.4. PROC SURVEYSELECT is designed to facilitate the task of sample selection, so it is generally more useful at the survey design stage rather than the analysis stage. Once all data has been collected, there are five additional SURVEY procedures available to produce descriptive statistics and conduct more sophisticated analytic tasks such as multivariate modeling.

Procedure	Analytic Tools
PROC SURVEYSELECT	Variety of built-in routines for selecting random samples; also contains methods for allocating a fixed sample size amongst strata and determining a necessary sample size given certain constraints
PROC SURVEYMEANS	Descriptive statistics such as means, totals, ratios, quantiles, as well as their corresponding measures of uncertainty
PROC SURVEYFREQ	Tabulations, measures of association, odds ratios, and risk statistics
PROC SURVEYREG	Regression models where the outcome is a continuous variable
PROC SURVEYLOGISTIC	Regression models where the outcome is a categorical variable
PROC SURVEYPHREG	Cox proportional hazards models for time-to-event data (survival analyses)

Table 1: Summary of SURVEY Procedures Available as of SAS Version 9.4

4. THE FOUR FEATURES OF COMPLEX SURVEYS

4.1 A HYPOTHETICAL EXPENDITURE SURVEY

To motivate exposition of the four features of complex survey data, suppose a market research firm has been hired to evaluate the spending habits of $N = 2,000$ adults living in a small town. Two example estimates of interest are the average amount of money an adult spent in the previous year on over-the-counter (OTC) medications and the average amount spent on travel outside the town. The discussion

below centers around a few (hypothetically carried out) sample designs to collect spending data for a sample of $n = 400$ adults and the specific complex survey features the designs introduce.

4.2 FINITE POPULATION CORRECTIONS

Suppose the first sample design involved compiling the names and contact information for all $N = 2,000$ people in the town onto a sampling frame and drawing a simple random sample (SRS) of $n = 400$ of them to follow up with to collect the expenditure information. From this sample, the estimated average of the

given expenditure y would be calculated as $\hat{\bar{y}} = \sum_{i=1}^{n=400} \frac{y_i}{n}$. The research firm might then reference an

introductory statistics textbook and calculate an estimated *element variance* of y as $s^2 = \sum_{i=1}^{n=400} \frac{(y_i - \hat{\bar{y}})^2}{n-1}$, an

unbiased, sample-based estimate of the population element variance, or $S^2 = \sum_{i=1}^{N=2000} \frac{(y_i - \bar{y})^2}{N-1}$, and use

this quantity to estimate the variance as of the sample mean by $\text{var}(\hat{\bar{y}}) = \frac{s^2}{n}$. They might then construct a

95% confidence interval as $\hat{\bar{y}} \pm 1.96\sqrt{\text{var}(\hat{\bar{y}})}$, where $\sqrt{\text{var}(\hat{\bar{y}})} = \text{se}(\hat{\bar{y}})$ is the standard error of $\hat{\bar{y}}$. Note how the standard error differs conceptually and computationally from the *standard deviation* of y , which is $S = \sqrt{S^2}$ for the full population and estimated by $s = \sqrt{s^2}$ from the sample.

It turns out the market research firm's calculations would overestimate the variance of the sample mean because, whenever the sampling fraction, n/N , is non-negligible (as is the case with $400/2,000$), there is an additional term that enters into variance estimate calculations called the *finite population correction*

(FPC). A more accurate estimate of the variance of the sample mean is $\text{var}(\hat{\bar{y}}) = \frac{s^2}{n} \left(1 - \frac{n}{N}\right)$, where the

term $\left(1 - \frac{n}{N}\right)$ is the FPC. Notice how the FPC tends to 0 as n approaches N . Since the goal of the

variance measure is to quantify the sample-to-sample variability of the estimate, an intuitive result is that it decreases as the portion of the population in each respective sample increases. In the extreme case when $n = N$, or when a census is undertaken, the variance accounting for the FPC would be 0. And despite the discussion in this section pertaining strictly to estimating the variance of a sample mean, a comparable variance formula modification translates to other statistics such as totals and regression coefficients.

The difference between the two variance perspectives is that the traditional formula implicitly assumes the data were collected under a simple random sampling with replacement (SRSWR) design, meaning each unit in the population could have been selected into the sample more than one time. Equivalently, the tacit assumption could be that the data were collected using simple random sampling without replacement (SRSWOR) from an effectively infinite population, a corollary of which is that the sampling fraction is negligible and can be ignored. Sampling products from an assembly line or trees in a large forest might fit reasonably well within this paradigm. But in contrast, survey research frequently involves sampling from a finite population, such as employees of a company or residents of a municipality, in which case adopting the SRSWOR design formulas is more appropriate.

There are two options available within the SURVEY procedures to account for the FPC. The first is to specify the population total N in the TOTAL= option of the PROC statement. The second is to specify the sampling fraction n/N in the RATE= option of the PROC statement. In terms of the sample design presently considered, specifying TOTAL=2000 or RATE=0.20 has the same effect. The syntax to account for the FPC is identical across all SURVEY procedures, and the same is true for the other three features of complex survey data as well.

Suppose the SAS data set SAMPLE_SRSWOR contains the results of this survey of $n = 400$ adults in the town. Program 1 consists of two PROC SURVEYMEANS runs on this data set. We are requesting the sample mean and its estimated variance for the over-the-counter expenditures variable (EXP_OTCMEDS). The first run assumes the sample was selected with replacement. Since there are no complex survey features specified, it produces the same figures that would be generated from PROC MEANS. The second requests the same statistics but specifies TOTAL=2000 in the PROC statement, in effect alerting SAS that sampling was done without-replacement and so an FPC should be incorporated. (The SURVEY procedure determines n from the input data set.)

Program 1: Illustration of the Effect of a Finite Population Correction on Measures of Variability

```

title1 'Simple Random Sampling without Replacement';
title2 'Estimating a Sample Mean and its Variance Ignoring the FPC';
proc surveymeans data=sample_SRSWOR mean var;
    var exp_OTCmeds;
run;

title2 'Estimating a Sample Mean and its Variance Accounting for the FPC';
proc surveymeans data=sample_SRSWOR total=2000 mean var;
    var exp_OTCmeds;
run;

```

Simple Random Sampling without Replacement
Estimating a Sample Mean and its Variance Ignoring the FPC

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	400

Statistics			
Variable	Mean	Std Error of Mean	Var of Mean
exp_OTCmeds	17.645854	0.683045	0.466550

Simple Random Sampling without Replacement
Estimating a Sample Mean and its Variance Accounting for the FPC

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	400

Statistics			
Variable	Mean	Std Error of Mean	Var of Mean
exp_OTCmeds	17.645854	0.610934	0.373240

The sample mean is equivalent (\$17.65) in both PROC SURVEYMEANS runs, but measures of variability are smaller with the FPC incorporated. Specifically, observe how the estimated variance of the mean has

been reduced by a factor of 20%, since $0.3732 = 0.4666 * \left(1 - \frac{400}{2000}\right)$. Since the standard error of the mean is just the square root of the variance, by comparison it has been reduced to $0.6109 = 0.6830 * \sqrt{1 - \frac{400}{2000}}$.

Where applicable, the FPC is beneficial to incorporate into measures of variability, because doing so results in increased precision and, therefore, more statistical power. There are occasions, however, when the FPC is known to exist yet intentionally ignored. This often occurs when effectively assuming a with-replacement sample design dramatically simplifies the variance estimation task (see discussion regarding the ultimate cluster assumption in Section 4.4), especially when there is only a marginal precision gain to be realized from adopting the without-replacement variance formula. Providing a few numbers to consider, with a sampling fraction of 10%, we would anticipate about a 5% reduction in the standard error; if the sampling fraction were 5%, the reduction would be around 3%. While the with-replacement assumption typically imposes an overestimation of variability, the rationale behind this practice is that the computational efficiencies outweigh the minor sacrifice in precision. The discussion in Rust et al. (2007) provides further insight into the trade-offs to be considered.

4.3 STRATIFICATION

The second feature of complex survey data is *stratification*, which involves partitioning the sampling frame into H mutually exclusive and exhaustive *strata* (singular: stratum), and then independently drawing a sample within each. There are numerous reasons the technique is used in practice, but a few examples are as follows:

- *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. In practice, sometimes the given subgroup's stratum is so small that it makes more sense to simply conduct a census of those units rather than select a sample of them.
- *Administer multiple modes of data collection.* To increase representation of the target population, some survey sponsors utilize more than one mode of data collection (de Leeuw, 2005). When the separate modes are pursued via separate sampling frames, these frames can sometimes be treated as strata of a more comprehensive sampling frame.
- *Increase precision of overall estimates.* When strata are constructed homogeneously with respect to the key outcome variable(s), there can be substantial precision gains achieved.

To illustrate how precision can be increased if the stratification scheme is carried out prudently, let us return to the expenditure survey example and consider an alternative sample design. Suppose there is a river evenly dividing the hypothetical town's population into an east and a west side, each with 1,000 adults, and that adults living on the west side of the river tend to be more affluent. It is plausible that spending behaviors differ markedly between adults on either side of the river. Since the two key outcome variables deal with expenditures, this could be a prudent choice for a stratification variable. Assume the firm is able to stratify their sampling frame accordingly, allocating the overall sample size of $n = 400$ adults into $n_1 = 200$ adults sampled without replacement from the west side and $n_2 = 200$ from the east.

Suppose the survey was administered and the results stored in a data set called `SAMPLE_STR_SRSWOR`.

Any time survey data emanate from a stratified random sampling design, we should specify the stratum identifier variable(s) on the survey data set in the `STRATA` statement of the `SURVEY` procedure. For our example here, we will assume the variable `CITYSIDE` defines which of the $H = 2$ strata the observation belongs to, a character variable with two possible values: "West" or "East."

Like Program 1, Program 2 consists of two `PROC SURVEYMEANS` runs on the survey data set, except this time we are analyzing a measure of travel expenditures (`EXP_TRAVEL`) instead of over-the-counter medications (`EXP_OTCMEDS`). The first run ignores the stratification and assumes a sample of size 400

was selected without replacement from the population of 2,000. The second run properly accounts for the stratification by placing CITYSIDE in the STRATA statement. Observe how the FPC is supplied by way of a secondary data set called TOTALS in the second run. This is because the FPC is a stratum-specific quantity. When there is no stratification or the stratification is ignored (as in the first run), one number is sufficient, but you can specify stratum-specific population totals (N_h 's) via a supplementary data set containing a like-name and like-formatted stratum variable(s) and the key variable _TOTAL_ (or _RATE_, if opting to provide sampling fractions instead).

Program 2: Illustration of the Effect of Stratification on Measures of Variability

```

title1 'Stratified Simple Random Sampling without Replacement';
title2 'Estimating a Sample Mean and its Variance Ignoring the
Stratification';
proc surveymeans data=sample_str_SRSWOR total=2000 mean var;
    var exp_travel;
run;

data totals;
    length cityside $4;
    input cityside _TOTAL_;
datalines;
East 1000
West 1000
;
run;

title2 'Estimating a Sample Mean and its Variance Accounting for the
Stratification';
proc surveymeans data=sample_str_SRSWOR total=totals mean var;
    strata cityside;
    var exp_travel;
run;

```

**Stratified Simple Random Sampling without Replacement
Estimating a Sample Mean and its Variance Ignoring the Stratification**

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	400

Statistics			
Variable	Mean	Std Error of Mean	Var of Mean
exp_travel	1363.179844	92.594306	8573.705490

Stratified Simple Random Sampling without Replacement
Estimating a Sample Mean and its Variance Accounting for the Stratification

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	2
Number of Observations	400

Statistics			
Variable	Mean	Std Error of Mean	Var of Mean
exp_travel	1363.179844	77.564901	6016.313916

The sample mean reported by PROC SURVEYMEANS is the same (\$1,363.18) in either case, but accounting for the stratification reduced the variance by almost one-third. Aside from a few rare circumstances, stratification increases the precision of overall estimates. It should be acknowledged, however, that any gains achievable are variable-specific and less pronounced for dichotomous variables (Kish, 1965). For instance, expenditures on over-the-counter medications are likely much less disparate across CITYSIDE, as expenditures of this sort seem less influenced by personal wealth than those related to travel.

Because sampling is performed independently within each stratum, we are able to essentially eliminate from consideration any between-stratum variability and focus only on the within-stratum variability. To see this, note how the estimated variance of the overall sample mean under this sample design is given by

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

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, SRSWOR sample means, where weights are determined by the proportion of the

population covered by the given stratum, or $\frac{N_h}{N}$, where $\sum_{h=1}^{H=2} \frac{N_h}{N} = 1$.

Figure 4 is provided to help visualize what is occurring in equation 1 when the SAMPLE_STR_SRSWOR data set is analyzed by PROC SURVEYMEANS with CITYSIDE specified in the STRATA statement. The vertical reference line represents the boundary between the two strata. For the west side ($h = 1$), the variance is a function of the sum of the squared distances between the points plotted and the horizontal reference line around \$2,500, the stratum-specific sample mean. For the east side ($h = 2$), the same can be said for the horizontal reference line around \$250. When the stratification is ignored, the vertical boundary disappears and a single horizontal reference line would replace the two stratum-specific lines around \$1,400 representing the combined mean expenditure across the two strata. The point is that the sum of the squared distances to this new reference line would be much greater, overall, which explains why measures of variability are larger in the second PROC SURVEYMEANS run when the stratification is ignored.

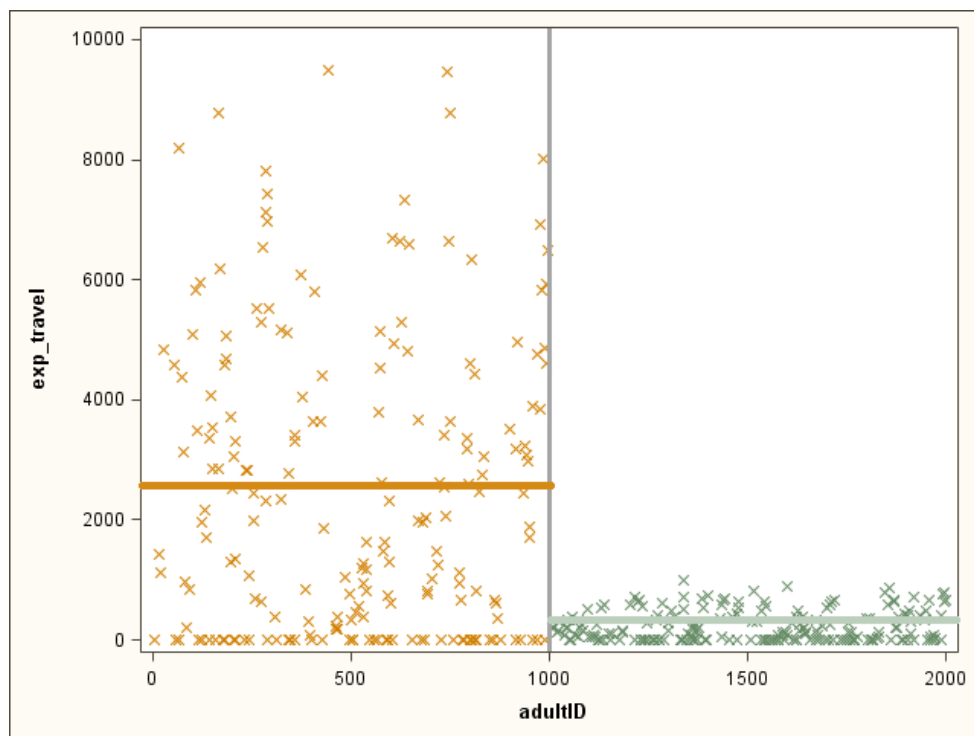


Figure 4: Visual Depiction of the Effect of Stratification on Variance Computations

4.4 CLUSTERING

The third feature of complex survey data is *clustering*. This occurs when there is not a one-to-one correspondence between sampling units and population elements; rather, the unit sampled is a cluster of population elements. A few examples include sampling households in a survey measuring attitudes of individuals, sampling doctor's offices in a survey measuring characteristics of patient visits to doctors' offices, or sampling classrooms in an education survey measuring the scholastic aptitude of students. Clustering is rarely ideal, as it generally decreases precision, yet it is often a logistical necessity or used to control data collection costs. For instance, most nationally representative, face-to-face surveys sample geographically clustered units to limit interviewer travel expenses.

Whereas homogeneity within strata fosters precision gains, homogeneity within clusters has the opposite effect. This phenomenon will be illustrated by considering the following alternative sample design for the expenditure survey example. Suppose the 2,000 residents are evenly distributed across the town's $C = 100$ blocks—that is, exactly $N_c = 20$ adults reside on each unique block—and that the market research firm decides data collection would be easier to orchestrate if the sampling units were blocks themselves. Perhaps they use a town map to enumerate all blocks and select a simple random sample of $c = 20$ of them, collecting expenditure data on all adults living therein. So the design still maintains a sample size of 400. Suppose the survey is administered and the results are stored in the data set called `SAMPLE_CLUS`. To isolate the effect of clustering, we will assume there was no stratification and, for simplicity, we will ignore the FPC.

Whenever the underlying sample design of the complex survey data set involves clustering, the cluster identifier variable(s) should be placed in the `CLUSTER` statement of the given `SURVEY` procedure. In the present example, this identifier is the variable `BLOCKID`. Program 3 is comprised of two `PROC SURVEYMEANS` runs, one assuming simple random sampling and another properly specifying the `BLOCKID` variable in the `CLUSTER` statement. As before, we are requesting the sample mean and its estimated variance, but this time for both expenditure variables, `EXP_OTCMEDS` and `EXP_TRAVEL`.

Program 3: Illustration of the Effect of Clustering on Measures of Variability

```
title1 'Cluster Sampling';
```

```

title2 'Estimating a Sample Mean and its Variance Ignoring the Clustering';
proc surveymeans data=sample_clus mean var;
  var exp_OTCmeds exp_travel;
run;

title2 'Estimating a Sample Mean and its Variance Accounting for the
Clustering';
proc surveymeans data=sample_clus mean var;
  cluster blockID;
  var exp_OTCmeds exp_travel;
run;

```

Cluster Sampling

Estimating a Sample Mean and its Variance Ignoring the Clustering

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	400

Statistics			
Variable	Mean	Std Error of Mean	Var of Mean
exp_OTCmeds	18.430203	0.709202	0.502968
exp_travel	1271.310549	101.074843	10216

Cluster Sampling

Estimating a Sample Mean and its Variance Accounting for the Clustering

The SURVEYMEANS Procedure

Data Summary	
Number of Clusters	20
Number of Observations	400

Statistics			
Variable	Mean	Std Error of Mean	Var of Mean
exp_OTCmeds	18.430203	0.814593	0.663563
exp_travel	1271.310549	320.315188	102602

This is yet another instance where ignoring a feature of the complex survey data set does not affect the point estimate, since the sample means are identical in both PROC SURVEYMEANS runs, but the clustering does impact measures of variability. In general, failing to account for clustering is especially hazardous because clustering can prompt a significant increase in the estimated variances. One might notice the increase is far more dramatic for EXP_TRAVEL than EXP_OTCMEDS. The explanation has to do with the degree of homogeneity, or how correlated adults' responses are within clusters, with respect to the given outcome variable.

The reader might find a visualization of homogeneity useful prior to the exposition of the equations commonly used to quantify it. Figure 5 below plots the distribution of the two expense variables within the sampled clusters. The cluster-specific mean is represented by a dot and flanked by a 95% confidence interval (not accounting for the any design features, only to illustrate the within-cluster variability). The idea is that the further away the dots are from one another, or the more dissimilar the confidence intervals appear, the larger the expected increase in variance when factoring in the clustering in the sample design. Contrasting the right panel to the left helps explain why the variance increase for travel expenditures trumps that for over-the-counter medications.

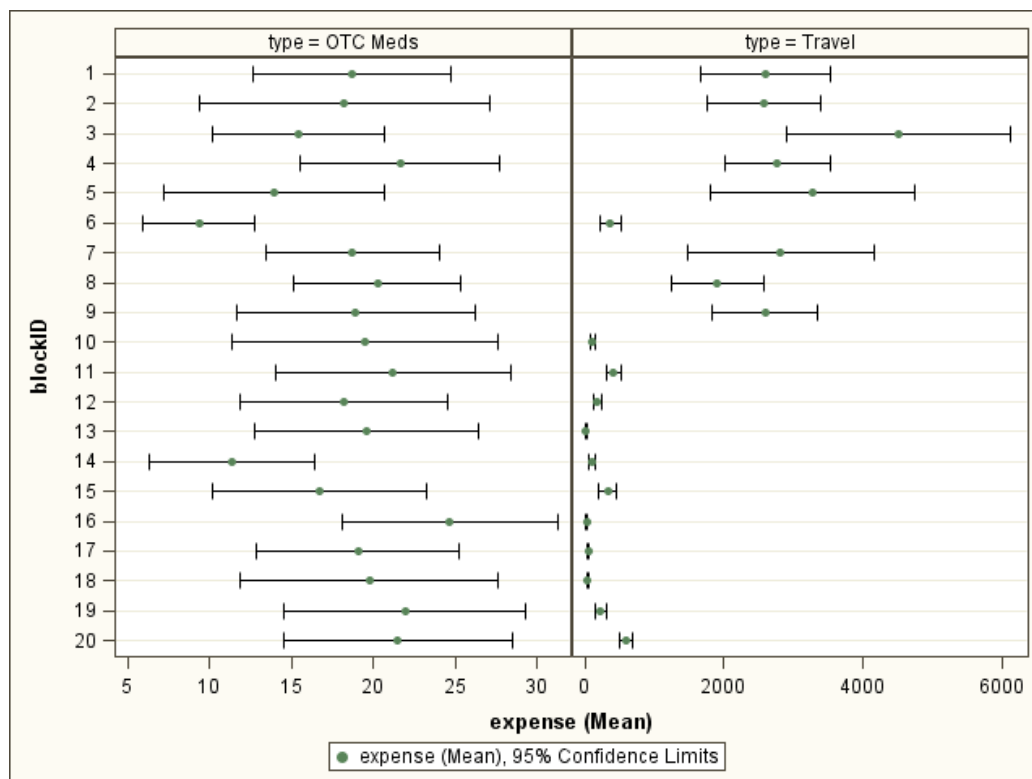


Figure 5: Expenditure Distributions within Blocks Selected as Part of a Clustered Sample Design for the Hypothetical Expenditure Survey

For the special case of equally-sized clusters, an alternative method to calculate the variance of a sample mean further elucidates this concept. Specifically, one can provide a summarized data set containing only the cluster-specific means to PROC SURVEYMEANS and treat as if it were a simple random sample. Program 4 demonstrates this method. It begins with a PROC MEANS step storing the cluster means of expenditure variables in a data set named CLUSTER_MEANS. The resulting data set of 20 observations is then analyzed by PROC SURVEYMEANS without any statements pointing to complex survey features.

Program 4: Illustration of the Effect of Clustering on Measures of Variability

```
proc means data=sample_clus noprint nway;
  class blockID;
  var exp_OTCmeds exp_travel;
  output out=cluster_means mean=;
run;

proc surveymeans data=cluster_means mean var;
  var exp_OTCmeds exp_travel;
run;
```

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	20

Statistics			
Variable	Mean	Std Error of Mean	Var of Mean
exp_OTCmeds	18.430203	0.814593	0.663563
exp_travel	1271.310549	320.315188	102602

Indeed, we can confirm the measures of variability output above match those from the second run of Program 3 in which we provided a data set of the full sample (400 observations) to PROC SURVEYMEANS but specified BLOCKID in the CLUSTER statement. The point of this exercise is to illustrate how clustering reduces the effective sample size. Kish (1965) coined the phrase *design effect* to describe this phenomenon.

The design effect of an estimate $\hat{\theta}$ is defined as the ratio of the variance accounting for the complex survey features to the variance under a simple random sample of the same size, or

$$Deff = \frac{\text{Var}_{\text{complex}}(\hat{\theta})}{\text{Var}_{\text{SRS}}(\hat{\theta})} \quad (2)$$

A *Deff* of 2 implies the complex survey variance is twice as large as that of a comparable simple random sample of the same size. Equivalently, this is to say the effective sample size is one-half the actual sample size. While it is possible for certain complex survey designs to render design effects less than 1, meaning designs that are more efficient than a simple random sample, clustering typically causes this ratio to be greater than 1.

In the special case of a simple random sample of equally-sized clusters, an alternative expression for equation 2 is

$$Deff = 1 + (N_c - 1)\rho \quad (3)$$

where N_c is the number of population elements in each cluster and ρ is the *intraclass correlation coefficient* (ICC), which measures the clusters' degree of homogeneity. A mnemonic device for remembering why the ICC is symbolized by a ρ is that represents the **rate of homogeneity**. This parameter is bounded by $-1/(N_c - 1)$ and 1. The extreme value on the lower end corresponds to all clusters sharing a common mean. A value of 1 at the other extreme indicates perfect homogeneity within clusters, or all elements therein sharing a common, unique value. In practice, negative values of ρ are rare. More common are slightly positive values; even if they seem small in absolute value at first glance, in the presence of large clusters, the variance increase could be substantial. That is, all else equal, larger clusters produce larger design effects. Lastly, note from equation 3 how clusters of size 1 would yield a design effect of 1, and thus default to an SRS variance estimate. (This is why clusters of size 1 on a survey data set input to a SURVEY procedure result in variances equivalent to those generated if the CLUSTER statement was omitted.)

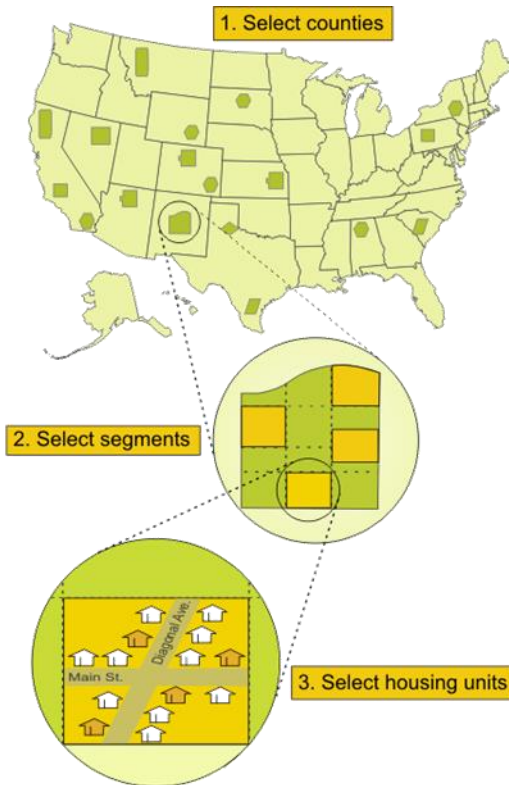
The true value of ρ is only known if we have knowledge about the entire population; however, a sample-based approximation is given by

$$\hat{\rho} = \frac{deff - 1}{N_c - 1} \quad (4)$$

where $deff$ is the sample-based estimate of $Deff$, or $deff = \frac{\text{var}_{complex}(\hat{\theta})}{\text{var}_{SRS}(\hat{\theta})}$.

There is no mandate to collect data on all population elements within a sampled cluster. For example, we could implement a *two-stage* sampling procedure in which a random sample of clusters is selected in the first stage, and then a subset of those elements therein is selected in the second stage. Returning to the current expenditure survey's clustered sample design, we could likely reduce the magnitude of the design effect while maintaining the same overall sample size if we altered the design such that we subsampled $n_c = 10$ adults within $c = 40$ sampled blocks as opposed to surveying all $N_c = 20$ adults within $c = 20$ sampled blocks, even though the net sample size is equivalent in either case. In the two-stage sample design, the blocks would be termed *primary sampling units* (PSUs) and adults *secondary sampling units* (SSUs). While increasing the number of PSUs in the sample design is desirable from a variance perspective, it may be offset by an associated increase in data collection costs. This is especially likely when the PSU is geographical in nature. Alas, there are trade-offs to consider when making fundamental sample design choices such as these.

Indeed, there may be even more than two stages of sampling. For example, Figure 6 depicts the multi-stage sample design employed by the Residential Energy Consumption Survey (RECS), a nationally-representative, interviewer-administered sample survey of U.S. households sponsored by the Energy Information Administration to collect data on energy usage and expenditures. The basic schema represented by Figure 6 is followed by countless other face-to-face surveys of U.S. households or the general U.S. population. The first step typically involves apportioning the land area of the U.S. into a set of PSUs consisting of single counties or groups of counties. A sample of these is then selected. Following that is a sequence of sampling stages for the various geographical units hierarchically nested within the PSUs, such as the census tracts, blocks, and households. For a practical discussion of area sampling, see Chapter 10 of Valliant et al. (2013).



Source: U.S. Energy Information Administration, 2009 Residential Energy Consumption Survey

Figure 6: Expenditure Distributions within Blocks Selected as Part of a Clustered Sample Design for the

Hypothetical Expenditure Survey

It is natural to assume each distinct stage of clustering in the sample design should be accounted for within the SURVEY procedure. Some competing software offers this capability, but SAS does not. The documentation repeatedly states users should only identify the PSU in the CLUSTER statement. This implicitly invokes the *ultimate cluster* assumption discussed in Kalton (1983, p. 35) and Heeringa et al. (2010, p. 67). In a maneuver to achieve computational efficiencies, the notion is to assume with-replacement sampling of the PSUs, in which case one only needs to focus on the variability amongst PSUs. Within-PSU variability, or any variability attributable to subsequent stages of sampling, is ignored.

Those unfamiliar with this concept are quick to express their apprehension about not explicitly accounting for all stages of sampling. Their initial reaction is that variances would be underestimated, since the terms associated with each stage are typically positive and additive. Most commonly, however, the with-replacement assumption leads to a slight *overestimation* of variability. For a few empirical studies demonstrating this tendency, see Hing et al. (2003) or West (2010).

To see why this occurs from an algebraic standpoint, consider equation 5, which represents the variance estimate for a sample mean expenditure under a design where c blocks are selected from the population of C blocks in the first stage, but only a sample of n_c of the N_c adults within each block are selected in the second stage (from p. 34 of Kalton, 1983):

$$\text{var}(\hat{y}) = \frac{s_c^2}{c} \left(1 - \frac{c}{C}\right) + \left(\frac{c}{C}\right) \frac{s_{n_c}^2}{cn_c} \left(1 - \frac{n_c}{N_c}\right) \quad (5)$$

where s_c^2 denotes the variance of PSU means amongst one another and $s_{n_c}^2$ is a term representing the average within-PSU variability. The with-replacement assumption is often employed when the first-stage sampling rate, c/C , is small, in which case the assumption is not far-fetched. Even if the without-replacement assumption were maintained in these instances, the first term of the summation tends to dominate, since the second term has this “small” fraction as a leading factor. Adopting the with-

replacement assumption of PSU selection eliminates all terms in equation 5 except $\frac{s_c^2}{c}$, and the crux of the Hing et al. (2003) and West (2010) case studies is that the variance increase associated with dropping $\left(1 - \frac{c}{C}\right)$ from the first term tends to outweigh the decreased associated with eliminating the second term entirely.

4.5 UNEQUAL WEIGHTS

The fourth feature of complex survey data is *unequal weights*. Unequal weights can arise for several reasons, but the reason we will focus on in this section is to compensate for unequal probabilities of selection. Up to this point in the paper, all example data sets have been derived from SRSWOR sample designs in which all sampling units have the same probability of being selected. Alternatively, variable selection probabilities can be assigned to the sampling units prior to implementing the randomized sampling mechanism, which can prove advantageous for a multitude of reasons.

In spite of certain benefits, variable selection probabilities complicate matters because we must introduce a weight w_i , which is referred to synonymously as a *sample weight*, *base weight*, or *design weight*, to compensate for the relative degrees of over-/under-representation of the particular units drawn into the sample. Specifically, Horvitz and Thompson (1952) showed that unbiased estimates of finite population quantities can be achieved by assigning each of the $i = 1, \dots, n$ units in the sample a weight of

$$w_i = \frac{1}{\Pr(i \in S)}, \text{ where } \Pr(i \in S) \text{ denotes the } i^{\text{th}} \text{ unit's selection probability into the given sample } S. \text{ Since the}$$

selection probabilities are strictly greater than zero and bounded by 1, sample weights are greater than or equal to 1. These can be interpreted as the number of units in the survey population (i.e., sampling frame) the sampling unit represents. For instance, a weight of 4 means the unit represents itself and three other comparable units that were not sampled. To have the SURVEY procedure properly account

for these weights during the estimation process, we simply specify the variable housing them in the WEIGHT statement.

Returning to the hypothetical expenditure survey, we will now motivate a sample design that renders unequal weights. When the sampling frame contains an auxiliary variable correlated with a key outcome variable to be measured in the survey, the technique of *probability proportional to size* (PPS) sampling can yield certain statistical efficiencies, particularly when estimating totals. Suppose the market research firm was able to obtain tax return information from the town's revenue office, information that included each of the $N = 2,000$ adults' income during the prior year. Despite the acquisition of this information being highly unlikely, it would make for an ideal candidate variable to use in a PPS sample design for an expenditure survey.

The notion behind PPS sampling is to assign selection probabilities in proportion to each sampling frame unit's share of the total for some *measure of size* variable. In general, if we denote the measure of size

for the i^{th} unit MOS_i and the total measure of size for all units in the sampling frame $\sum_{i=1}^N MOS_i$, the

selection probability for this unit would be $\Pr(i \in S) = n \times \left(\frac{MOS_i}{\sum_{i=1}^N MOS_i} \right)$ and, if the unit is selected into the

sample, its sample weight would be assigned as the inverse of that quantity.

Suppose the market research firm drew a PPS sample of $n = 400$ adults using income as the measure of size, carried out the survey, and the results have been stored in the data set SAMPLE_WEIGHTS. The base weights are maintained by the variable WEIGHT_BASE. The two PROC SURVEYMEANS runs in Program 5 both request the mean and variance of the two expenditure variables, but only the second run properly accounts for the weights (for simplicity, the FPC is ignored).

Program 5: Illustration of the Effect of Unequal Weights

```

title1 'Sampling with Weights';
title2 'Estimating Two Sample Means and Variances Ignoring the Weights';
proc surveymeans data=sample_weights mean var;
    var exp_OTCmeds exp_travel;
run;

title2 'Estimating Two Sample Means and Variances Accounting for the
Weights';
proc surveymeans data=sample_weights mean var;
    var exp_OTCmeds exp_travel;
    weight weight;
run;

```

Sampling with Weights
Estimating Two Sample Means and Variances Ignoring the Weights

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	400

Statistics			
Variable	Mean	Std Error of Mean	Var of Mean
exp_OTCmeds	18.042657	0.686190	0.470857
exp_travel	3771.900098	146.547307	21476

Sampling with Weights
Estimating Two Sample Means and Variances Accounting for the Weights

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	400
Sum of Weights	1715.42534

Statistics			
Variable	Mean	Std Error of Mean	Var of Mean
exp_OTCmeds	17.484794	0.857940	0.736061
exp_travel	1884.486868	124.739413	15560

Notice how this time both the sample means *and* measures of variability differ between the two runs. Recall that in the three previous illustrations of failing to account for a particular complex survey feature, only the measures of variability differed. This is because the previous three employed equal probability of selection designs, a byproduct of which is identical sample weights for all units in the sample. Although we could have computed and specified these weights in the WEIGHT statement, it would be technically

unnecessary, because when the weights are all equal, the weighted mean $\hat{\bar{y}}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$ is equivalent to

the unweighted mean $\hat{\bar{y}}_{unw} = \frac{\sum_{i=1}^n y_i}{n}$, since $\hat{\bar{y}}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} = \frac{w \sum_{i=1}^n y_i}{wn} = \frac{\sum_{i=1}^n y_i}{n}$. In effect, omitting the WEIGHT

statement is tantamount to assigning $w_i = 1$ for all units in the sample.

Observe how the difference between the unweighted and weighted sample means for EXP_TRAVEL is much more pronounced than the difference for EXP_OTCMEDS. Kish (1992) showed that the difference in these two estimators can be expressed as

$$\hat{y}_{unw} - \hat{y}_w \approx \frac{-\text{cov}(w_i, y_i)}{\bar{w}} \quad (6)$$

where $\text{cov}(w_i, y_i)$ is the covariance between weight variable w_i and the outcome variable y_i , and \bar{w} is the average weight in the sample data set. Equation 6 is useful for contemplating or explaining the potential discrepancies that can surface when failing to account for unequal weights when the estimate of interest is the mean. It implies there must be some kind of association between the weights and outcome variable for a difference to occur. In the current PPS sample design, there is an over-representation in the sample of higher-income adults, and it is evident they spend more on travel; the converse is true for lower-income adults. Using the weights rebalances these respective cohorts to better match their prevalence in the underlying population. On the other hand, the relationship between over-the-counter medications and the weights is clearly much weaker, as there is hardly any difference between two estimators. This serves to remind us how the impact is variable-specific. It appears the case for this particular outcome variable that $\text{cov}(w_i, y_i) \approx 0$.

4.6 BRIEF SUMMARY OF THE FEATURES OF COMPLEX SURVEY DATA

Table 2 summarizes the syntax required to account for all four features of complex survey data. The syntax is common across all SURVEY analysis procedures. Although the discussion above illustrated each feature in isolation to focus attention on its impact on estimates and/or measures of variability, a given complex survey data set encountered in practice can be characterized by any combination of these four features. It would be excessive to motivate and discuss a sample design utilizing any and all possible combinations. Based on this author's experiences, the FPC is the least commonly encountered feature, particularly for surveys that involve clustering—whether in single or multiple stages—because these surveys are liable to adopt the ultimate cluster (with-replacement) assumption of PSUs, in which case the first-stage FPC would be ignored anyhow. Perhaps the most commonly encountered combination is stratification, clustering, and unequal weights.

Feature	Syntax	Comments/Warnings
Finite Population Correction	TOTAL=(value(s)) data-set-name or RATE=(value(s)) data-set-name in the PROC Statement	When sample design is stratified, you can specify a supplementary dataset with stratum identifier variable(s) and their corresponding stratum-specific population total (_TOTAL_) or rate (_RATE_). Stratum variable(s) in supplementary data set must be equivalently named and formatted.
Stratification	STRATA statement	Stratum identifier variable(s).
Clustering	CLUSTER statement	Cluster identifier variable(s). Even if sampling involves multiple stages of clustering, specify only the primary sampling unit (PSU), or the first-stage identifier.
Unequal Weights	WEIGHT statement	Only observations in the input data set with positive weights ($w_i > 0$) are maintained. Without a WEIGHT statement, an implicit weight of 1 is applied to all observations.

Table 2: Summary of SURVEY Procedure Syntax to Account for the Four Complex Survey Data Features

5. SUMMARY

This paper began by laying out some of the terminology and issues involved when administering a survey. A hypothetical expenditure survey being conducted by a market research firm was used to motivate some of the decisions that must be made regarding the sample design and data collection

methods, as well as which the particular complex survey features these decisions introduce. The main takeaway message of this paper is that complex survey features alter most of the familiar formulas taught in introductory statistics courses. Some alterations lead to decreased measures of variability, some to increased measures of variability, while others lead to a different point estimate. When analyzing a data set derived from a complex survey, consult any resources available (e.g., subject matter experts, technical reports, or other forms of documentation) to determine whether any of the features discussed in this paper are present. If so, you should use one of the SURVEY analysis procedures listed in Table 1 to formulate estimates from it.

REFERENCES

- Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R. (2009). *Survey Methodology*. Hoboken, NJ: Wiley.
- Heeringa, S., West, B., and Berglund, P. (2010). *Applied Survey Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press.
- Hing, E., Gousen, S., Shimizu, I., and Burt, C. (2003). "Guide to Using Masked Design Variables to Estimate Standard Errors in Public Use Files of the National Ambulatory Medical Care Survey and the National Hospital Ambulatory Medical Care Survey," *Inquiry*, **40**, pp. 401 – 415.
- Holt, D., and Smith, T. (1979). "Poststratification," *Journal of the Royal Statistical Society, Series A*, **142**, pp. 33 – 46.
- Horvitz, D., and Thompson, D. (1952). "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association*, **47**, pp. 663 – 685.
- Izrael, D., Hoaglin, D., and Battaglia, M. (2000). "A SAS Macro for Balancing a Weighted Sample," *Proceedings of the SAS Users Group International (SUGI) Conference*. Available on-line at: <http://www2.sas.com/proceedings/sugi25/25/st/25p258.pdf>
- Kalton, G. (1983). *Introduction to Survey Sampling*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-035. Newbury Park, CA: Sage.
- Kish, L. (1965). *Survey Sampling*. New York, NY: Wiley.
- Kish, L. (1992). "Weighting for Unequal P_i ," *Journal of Official Statistics*, **8**, pp. 183 – 200.
- de Leeuw, E. (2005). "To Mix or Not to Mix Data Collection Modes in Surveys," *Journal of Official Statistics*, **21**, pp. 233 – 255.
- Lepkowski, J., Mosher, W., Davis, K., Groves, R., and Van Hoewyk, J. (2010). *The 2006–2010 National Survey of Family Growth: Sample Design and Analysis of a Continuous Survey*. *Vital Health Statistics, Series 2*, **150**. Hyattsville, MD: National Center for Health Statistics.
- Lohr, S., and Rao, J.N.K. (2000). "Inference from Dual Frame Surveys," *Journal of the American Statistical Association*, **95**, pp. 271 – 290.
- Lohr, S. (2010). *Sampling: Design and Analysis. Second Edition*. Boston, MA: Brooks/Cole.
- Rust, K., Graubard, B., Kott, P., and Eltinge, J. (2007). Washington Statistical Society President's Invited Panel Discussion on Finite Population Correction Factors, Bureau of Labor Statistics, Washington, DC, March 27. Slides available on-line at: <http://washingtonstatisticalsociety.org/seminars/sem2007.html>
- Valliant, R., Dever, J., and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York, NY: Springer.
- West, B. (2010). "Accounting for Multi-Stage Sample Designs in Complex Sample Variance Estimation," *Institute for Social Research Technical Report Prepared for National Survey of Family Growth (NSFG) User Documentation*. Available on-line at: http://www.isr.umich.edu/src/smp/asda/first_stage_ve_new.pdf

RECOMMENDED READING

Complex Survey Data Analysis in SAS® (forthcoming) by Taylor Lewis

CONTACT INFORMATION

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

Taylor Lewis
PhD Graduate
Joint Program in Survey Methodology
University of Maryland, College Park
tlewis9@umd.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.