The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata and Mata are registered trademarks of StataCorp LP.

The Stata Journal (2007)
7, Number 1, pp. 1–21

A survey on survey statistics: What is done and can be done in Stata

Frauke Kreuter
Joint Program in Survey Methodology
University of Maryland, College Park
fkreuter@survey.umd.edu

Richard Valliant
Joint Program in Survey Methodology
University of Michigan, Ann Arbor

Abstract. This article will survey issues in analyzing complex survey data and describe some of the capabilities of Stata for such analyses. We will briefly review key elements of survey design and explain the effects of different design features on bias and variance. We compare different methods of variance estimation for stratified and clustered samples and discuss the handling of survey weights. We will also give examples for the practical importance of Stata's survey capabilities.

Keywords: st0118, cluster sampling, complex design, nonresponse, stratified sampling, variance estimation, weights, DEFECT, NHANES, NHIS, PISA

1 Issues in analyzing survey data

Survey data are used in most empirical work in behavioral and social sciences, economics, and public health. Throughout the last few years, there has been an increased awareness that researchers need to consider the sampling design when analyzing survey data. The increasing awareness led several of the major statistical software packages to expand their features for analyzing complex survey data. Survey statisticians recognize Stata as one of the most powerful packages. However, applied substantive researchers still do not always account for survey design information as part of their standard practice. This article will therefore provide a rough guideline through the various Stata methods that are appropriate for analyzing survey data and should help to answer the following questions:

- What are the survey design features that I need to take into account?
- Why do I need to take these survey design features into account? How do such survey features affect bias and variance?
- How do I account for complex designs in practice?

This article's goal is not to explain all possible survey designs but rather to fill some knowledge gaps about issues that need to be considered in day-to-day data analysis. We will start with a brief review of the common elements of complex survey designs in section 2 and discuss the consequences of excluding these elements in section 3. Readers who are already familiar with sampling designs may skim these sections and continue with section 4, where we discuss two major variance estimation methods for complex

surveys: Taylor linearization and replication. In section 5, we demonstrate the use of Stata procedures in analyzing public-use data for two large-scale surveys. The article concludes with a brief summary.

2 Features of survey design

Estimates produced by standard procedures in statistical packages usually ignore survey design features and assume that observed data are realized values of independent random variables or that the data were collected from a simple random sample (SRS). In contrast, sample surveys involve three features that have potentially significant consequences for estimation: weights, stratification, and clustering. We will briefly introduce these features before we discuss their effects and related problems.

Another feature of many surveys is that, in practice, sampling is typically done without replacement to avoid multiple selections of the same sampling unit. The resulting difference in variance estimates for with- and without-replacement samples is negligible if the sample is a small proportion of the population. Because this proportion is small in our examples, as it is for many survey data, we will not discuss this issue further.

Most surveys begin with a probability sample from a population frame. When the population is relatively small, the frame may be a list of all units in the population. For example, if a survey is conducted of all elementary schools in a region, a list may be available from a government education agency. In countries with population registries, those might be used as a sampling frame for household surveys. Sometimes the frame may not fully cover the desired population, but the weighting step, described below, tries to correct for this.

Weights: Survey weights are designed to expand the sample to the level of the population that the sample represents. In a probability sample, units are selected using known probabilities. In some surveys, all units have the same selection probability, but more typically there will be some variation in the probabilities. In a survey of persons, separate analyses of groups defined by age, gender, and race–ethnicity may be planned. Consequently, those groups may be sampled at different rates to obtain adequate sample sizes from each. The selection probabilities account for unequal sampling rates used for different types of units. The inverse of the selection probability of a sample unit is known as its *base weight*. For example, if males were selected with probability 0.01 and females with probability 0.05, the base weights for males and females would be 100 and 20, respectively.

Many survey datasets are delivered with what are called *final weights* that not only take sampling probabilities into account but are also designed to adjust for nonresponse, coverage problems, and other uses of auxiliary data outside the survey.

Stratification: With *stratification*, population elements are divided into strata: mutually exclusive and exhaustive subgroups. That is, some information for every element

needs to be on the frame of population elements to divide them into strata. For example, telephone numbers for surveys of U.S. households are often divided into geographical strata. To do so, the researcher must be able to identify the geographic region of each telephone number in the sampling frame. The left panel in figure 1 shows a population that is divided into five strata (indicated by solid lines). Sampling then takes place within each of these strata. The x's in the left panel of figure 1 denote four selected sample units in each of these five strata. One reason to stratify is the desire to make comparisons among the subgroups that form the strata, and stratification ensures that units from each group are selected into the sample. Political or geographical regions are often used as strata for this reason.

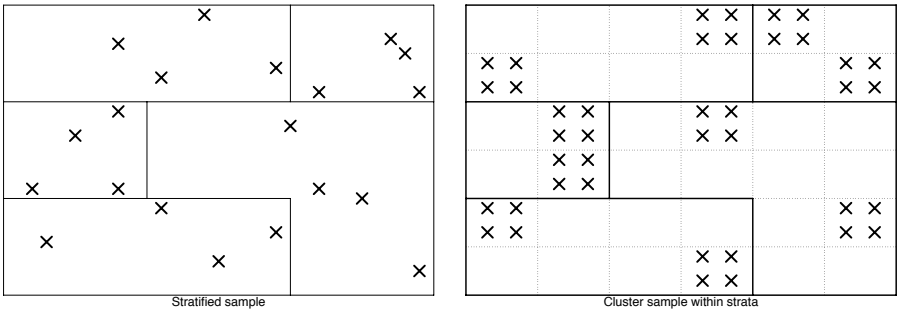

Figure 1: Stratified and clustered samples

Clustering: Samples are called *clustered* if one specifies groups of population units, and a sample of such groups (*primary sampling units* [PSUs]) is first taken instead of the individual units. The dotted lines in the right panel of figure 1 indicate such clusters within the strata. Here two PSUs are selected in each of the five strata. In this simple example of cluster sampling, all elements within each cluster are selected into the sample. Researchers often decide to use a clustered sample instead of a simpler design for organizational or financial reasons. The absence of a general population registry in many countries makes in-person surveys of an SRS virtually impossible. Sampling in several stages, one of them at the level of small geographical clusters, facilitates selecting respondents without the aid of registry data. This approach is used in many household surveys when data are collected by in-person interviews and a list of all households is not available. Here geographic areas are sampled until, at the last stage, households can be listed and sampled. More complex designs can have further sampling within the clusters. Also, a sample in which geographical clusters are sampled first is cost efficient for face-to-face surveys, since interviewing respondents who live close together reduces travel costs.

3 Accounting for survey design: Effects on bias and variance

Two challenges arise when dealing with survey data: (1) obtaining correct point estimates (avoiding bias) and (2) computing correct variances and standard errors (SEs). The three elements described above (weights, stratification, and clustering) have different effects on bias and variance.

3.1 Weights

If the sample is selected with *unequal selection probabilities*, disregarding sampling weights can lead to biased estimates when estimating population totals, means, or other more complicated quantities. If weights are used in models, the resulting estimates are of models that would be fitted if you had the entire population in the sample. But even if a sample is selected with equal selection probabilities, analysts might be confronted with weights in the resulting dataset. Those weights are usually designed to adjust for nonresponse or coverage error (or both). Typically users will not create those weights themselves. Datasets are usually delivered with weight variables designed by the data producer.

Most complex samples suffer some degree of nonresponse. Nonresponse can occur for several reasons. For example, in a household survey, contact may never be made with some households because no one can be found at home during the survey period. Others that are contacted may refuse to participate. Only if the respondents can be safely treated as a random subsample of the full sample will estimates of quantities like means and proportions be unbiased. Nonresponse can lead to bias if the response mechanism is related to the outcome variable (Groves et al. 2004). For example, if older persons are less likely to respond in a health survey than younger persons and are more likely to be sick, then an estimate of the proportion of sick persons could be too low. A standard method to compensate for the units lost to nonresponse is to classify all sample units (responders and nonresponders) into cells on the basis of characteristics that are predictors of whether a unit does respond. In the example above, age would be such a characteristic. The responders will then get a weight assigned that compensates for the missing cases in each cell. This method requires not only knowledge about relevant characteristics but also that responders in a cell can be treated as a random sample of the initial sample, e.g., that each unit within a given age group has the same probability of responding. Other more elaborate techniques using propensity scores (see Little and Rubin 2002, sec. 3.3) are available.

To correct for undercoverage of the target population, researchers often use auxiliary or predictor variables to *poststratify* the data. This method is commonly used in household surveys and involves adjusting the weights of respondents to force them to sum to population counts for different groups. For example, most U.S. household surveys do not adequately cover certain demographic groups. In the Current Population Survey, survey estimates of the number of young black males are only about 3/4 of the census

counts prior to poststratification (Kostanich and Dippo 2002). Poststrata might then be defined by age group crossed with gender and race. The weights of respondents in the poststratum would be adjusted to sum to the count in that group from the most recent census. In the Current Population Survey, the poststratification adjustment increases the weights of the young, black male responders by about 4/3. Similar to nonresponse adjustment, poststratification relies on the assumption that noncovered persons can be treated as missing at random within each poststratum.

When nonresponse adjustment and poststratification are both used, the final survey weight for a responding unit j has the form $w_j f_{NRj} f_{PSj}$, where w_j is the base weight and f_{NRj} and f_{PSj} are the nonresponse and poststratification adjustments applied to unit j. The final weight appears on each respondent data record as a variable to be used in the analysis of the data. Some datasets provide both the individual components (base weights, nonresponse weights, poststratum weights) and the final product (Groves et al. 2004), but having only the final weight is probably more common. This final weight is all that is necessary for most analyses.

If the assumptions are met about why data are missing, applying weights can reduce bias in the estimates of means, proportions, totals, etc. At the same time, SEs can increase because of the use of weights. The increase in SEs is sometimes used as an argument against weights. However, excluding weights gives estimates that may apply only for the sample and not to the full population.

3.2 Stratification

Dividing the frame of population elements into strata to ensure the possibility of making comparisons across those strata (e.g., geographical regions) is only one reason for stratification. Another reason is the reduction in sampling variation that one can achieve with stratification: the variation from sample to sample is restricted to the variation within strata. Often strata reflect groups that are more homogeneous than the population as a whole. Here a sample drawn with stratification and an efficient allocation to strata will lead to smaller SEs for estimation of population statistics than those of a sample without stratification. The effect of stratification on point estimates will be reflected in how the weights are calculated. Suppose that one were to use geographic strata and the survey variable of interest is skin cancer. Assume that the prevalence of skin cancer varies with environmental exposure in different geographic regions. Here the stratified sample would ensure that elements from each region would be selected for the sample. The sample-to-sample variation in environmental exposure is limited to the variation within region.

Consequently, if the data are collected using a stratified sampling design, analysts should take the stratification information into account when estimating SEs; otherwise, the resulting SEs will be incorrect. In stratified samples that do not involve clustering, the SEs will often be too large if the strata are ignored. This is especially true when the strata form homogeneous groups. If the strata are not particularly homogeneous, accounting for them still produces approximately unbiased SEs. From an analyst's point of view, the potential decrease in SEs is an incentive not to neglect the sampling design.

3.3 Clustering

Whereas having homogeneous groups is an advantage for stratification where elements are taken from each stratum, it is a disadvantage for cluster samples where only some clusters get selected for any given sample. In practice, units in clusters used for household surveys are often near each other to save interviewer travel costs. And people who live close together are likely to be similar in their economic background, educational level, or accessible infrastructure.

If clusters are more homogeneous than the population, estimates from cluster samples will have larger SEs than estimates from an SRS of equal size for two reasons. For one, if only a subset of these homogeneous clusters is selected, the resulting data might vary from sample to sample more than they would if an SRS (or another type of single-stage element sample) had been used. Second, the similarity within the clusters can be equated to a decrease in sample size. In an extreme case, where all elements of the cluster have the same value on the variable of interest, the sample size will in effect be reduced to the number of clusters. For example, in figure 2 two clusters are selected within each stratum. The elements within each cluster are all equal, and they are different from those in the second cluster in each stratum. The eight elements selected in the two clusters within each stratum carry only as much information as one would have gotten with one element from each cluster. The effective sample size in each stratum is reduced from 8 to 2 observations. A decrease in sample size will, however, increase variances. Also, the cluster-induced similarity will violate the standard assumption of having independent observations. Although failing to take the clustering information into account will still provide correct point estimates, SEs of many statistics will be underestimated, which means that results will falsely appear to be statistically significant.

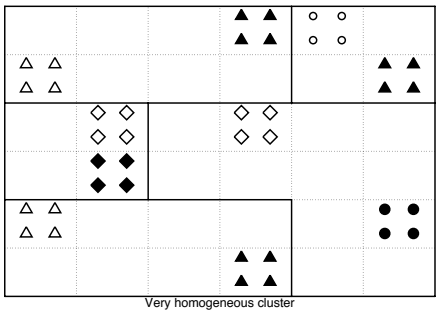

Figure 2: Sample with homogeneous clusters

Researchers should think of clustering effects in their own area of expertise. Clusters can be students nested within teachers, patients within doctors, or employees within businesses. For survey data, we discussed effects of the relative homogeneity of respondents within the same geographical area. However, even in samples that are not geographically clustered, such as random-digit-dialed telephone surveys, different respondents' answers can still be correlated (Groves 1989, 318). Data from the design effect study (DEFECT) in Germany (Schnell and Kreuter 2000) show the added cluster effect that is due to interviewers in face-to-face interviews (Schnell and Kreuter 2005). Here too it appears that for certain items the homogenizing effect of interviewers is even higher than the homogenizing effect of the geographical area.

3.4 Examples

NHANES: The estimates in table 1 show a potential bias-inducing effect of complex samples. The point estimate of the percentage of people suffering from hypertension would have been 5.4% by using unweighted data from the National Health and Nutrition Examination Survey (NHANES) III, Phase 2. However, the NHANES is a geographically stratified, multistage, clustered sample of households with age, sex, and race–ethnic groups sampled at different rates. The design includes 23 strata with two sample PSUs per stratum.

The estimate for hypertension when all design information is ignored is 5.4%. The estimated SE is 0.25, again ignoring all design features. If stratification and clustering are taken into account, the SE increases to 0.34. The point estimate for the percentage of people suffering from hypertension is still 5.4%. The point estimate does change as soon as the weights are applied. In particular, the estimate for the percentage suffering from hypertension decreases to 3.9% (see rightmost column in table 1). The decrease occurs because groups with higher rates of hypertension are also oversampled in NHANES. Those groups thus have smaller weights than groups with lower incidence, causing the weighted proportion to be substantially less than the unweighted proportion. Taking the weights into account has an additional effect on the estimated SEs.

Table 1: Percentage of hypertension among 8,344 adult respondents in NHANES III

	Ignoring all design information	Accounting for stratification and clustering	
		Unweighted	Weighted
Hypertension, %	5.4	5.4	3.9
SE	0.25	0.34	0.43

Taking all design information into account (stratification, clustering, and design weights), the SE is now 0.43. The ratio of 0.43^2 (variance accounting for complexity) to 0.25^2 (variance for misspecified model) is called the misspecification effect (*meff*). Here the variance is misspecified by 2.96 if all survey design and estimation features are

ignored when doing an analysis and when the sample is treated as if it had been selected by an SRS with replacement. The correctly estimated SE is thus about $\sqrt{2.96} = 1.7$ times as large as the incorrectly estimated SE. The intermediate estimate of 0.34, although larger than 0.25, is still not an acceptable measure of error because it does not account for the bias of the unweighted point estimate, which is substantial in NHANES.

In a more complicated analysis than the one above, an analyst might run an ordinary least squares regression on a clustered household sample where Hispanics were sampled at a much higher rate than non-Hispanics. Using ordinary least squares would entail at least two types of misspecification. First, the proportion of Hispanics in the sample would be much higher than in the population; ignoring the weights fails to correct for this imbalance. Second, ordinary least squares ignores clustering, which will typically lead to an underestimation of SEs of model parameter estimates.

Next to meff there is a second measure called *deff* (Kish 1965) used in survey statistics. The deff is defined as the ratio of the variance accounting for complexity over the variance if SRS had been used to select the sample. If the sampling weights are the same for all elements in the sample, deff and meff take on equal values. The deff is used for survey design planning, particularly in determining sample sizes and costs. The deff is also sometimes reported with survey results, especially if variables necessary to correct SEs are not provided with the data files. Effects on SEs are usually reported as deft $= \sqrt{\text{deff}}$ (Kish 1965).

PISA: The sometimes drastic effect of ignoring sampling design information is displayed in figure 3, which shows the confidence intervals for the average reading scores in Denmark compared with the United States by using data from the Organisation for Economic Co-operation and Development (OECD)–sponsored Programme for International Student Assessment (PISA) in 2000. In PISA, a sample of schools and students within those schools was selected to measure reading, mathematics, and science literacy in 32 countries.

The point estimates between the two countries differ by seven points with an average score of 496.56 for Denmark and 503.71 for the United States. A naive test of mean differences ignoring the design information would have led to the false impression that there is a significant difference between the reading scores of these two countries [$F(1, 8080) = 7.75$ and a p-value of 0.0054]. In figure 3 for Denmark, the size of the SEs does not change too much when accounting for the complex design, and the confidence intervals are only slightly wider. However, for the United States, the confidence intervals are vastly underestimated when SEs are computed as if these data come from an SRS. When we account for the complex design, the difference between United States and Denmark is no longer significant [$F(1, 79) = 0.93$ and a p-value of 0.3380; see p. 18].

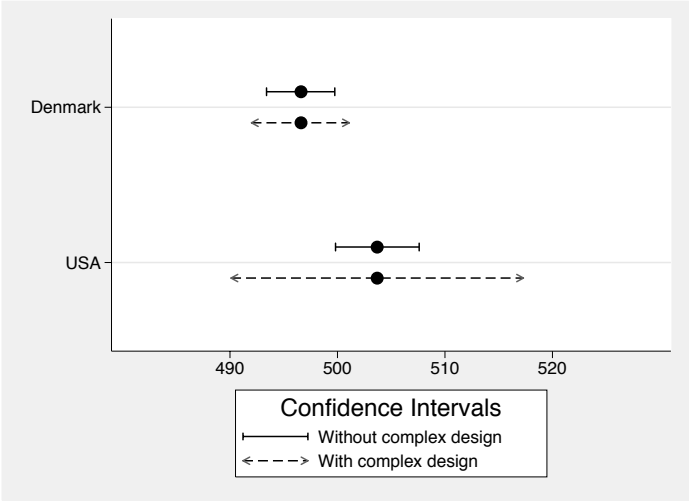

Figure 3: Differences in reading scores between Denmark and the United States. Solid lines indicate the confidence intervals around the weighted means but ignoring stratification and cluster information. Dashed lines reflect the confidence intervals after computing the correct SEs.

4 Variance estimation

Analysts have used three basic strategies to account for complex designs when estimating SEs of descriptive statistics or model parameters. The first is to simply multiply SEs from standard analyses by an (outside) estimate of a deft. The second approach, used when fitting a model, is to include terms that implicitly account for design features. The third strategy is to use software that directly estimates SEs that account for the complex design. We discuss each of these below. Which method to use depends on the level of design complexity and on which variables are provided with the data to account for that complexity. One can do little if no weights and no design variables are provided.

The first strategy, mentioned above, is to run an analysis using standard software that ignores the sample design and then to adjust the SEs by a deft. The initial analysis may or may not incorporate survey weights. Defts can come from published values as they might appear in data descriptions or documentation or from using rules of thumb like deft = 1.4 (Kostanich and Dippo 2002). A rule of thumb should be based on surveys that are similar to the one you are dealing with. However, this approach is usually too crude because defts vary depending on all the factors discussed in section 2 and on the particular analysis being done.

The second approach, when fitting a model, is to include terms that implicitly incorporate design features. For example, dummy variables for strata may be included as additional independent variables in the model. This method is also unlikely to ade-

quately account for design features. This is especially true in cluster samples if nothing is done to account for correlation among units within clusters.

The first two approaches may be used when only partial design information is available. However, both are largely historical techniques that were used when software and computing power limited an analyst's ability to compute correct SEs.

The third, and best, approach is to use software packages that allow the estimation of SEs with methods that account for complex designs. The two general methods of doing this are linearization and replication. Estimating SEs for survey data does require information about the survey design to be part of the dataset. Information on stratification and clustering is provided in two forms in datasets. The data must either contain variables that indicate the design strata and design cluster or contain a set of replicate weights. If the design variables (strata and cluster) themselves are available, exact formulas or linearization (also known as Taylor series) estimation can be used. Exact formulas can be used for simple estimators, like totals, from basic designs like stratified SRS. The exact formulas are special cases of the more general linearization estimators. Linearization is needed for more complicated estimators, even when the design itself is simple. If replicate weights are provided in the data, some form of replication method will be used to estimate the correct SEs. Like linearization, replication can be used for complicated estimators even if the design is simple.[1]

Survey datasets can be set up to use either linearization or replication, and Stata supports both. Stata will also allow creating replication weights by using design variables. In sections 4.1 and 4.2, we briefly sketch the mechanics behind the two techniques.

The rest of this section will also list some of the relative advantages and disadvantages of the linearization and replication methods and summarizes some of the reasons why users may want both options to be available. A summary of pros (+) and cons (−) is given in table 2. Both approaches are good for most things, and the information provided with the dataset often dictates the choice.

1. The complications when estimating variances derive from the fact that estimators are often nonlinear. This nonlinearity is not unique to surveys, and the options for estimating variances for nonlinear estimators are the same as in the rest of statistics. With survey data even a simple statistic like a mean can be nonlinear since means are estimated as a weighted sum of data divided by a sum of weights. Since the denominator of this type of ratio mean is random in many sample designs, the mean itself is a ratio of random quantities and therefore nonlinear. Nonresponse adjustments and poststratification, mentioned earlier, also lead to nonlinear estimates.

Table 2: Comparing Taylor linearization and replication variance estimation

Feature	Linearization	Replication
Good large-sample properties	+	+
Maximizing degrees of freedom (stability)	+	–
Model robustness	+	+
Reflection of adjustments (e.g., poststratification)	–	+
Accounting for WOR sampling in multiple stages	+	–
Applicability to complex forms of estimates	+	+
Computational speed	+	–
Design knowledge needed	–	+
Separate variance formulas for different estimates	–	+
Disclosure of PSUs and strata	–	+
File size	+	–

NOTE: WOR = without replacement; PSU = primary sampling unit.

4.1 Linearization

Linearization, also known as Taylor series estimation or the delta method, involves making a linear approximation to the nonlinear statistic being analyzed. A variance formula, appropriate to the sample design, is then applied to that approximation. Stata and other statistical software have programmed the approximations for many statistics and require the user to specify only the analysis and certain information about the sample design. The theory justifying the method requires that many first-stage units, i.e., PSUs, be sampled (Krewski and Rao 1981). Thus designs with either a limited number of strata and many PSUs per stratum or few PSUs per stratum but many strata both qualify. Possibly the most general formulation of linearization estimators for complex surveys is from Binder (1983). The sandwich estimator discussed by Binder (1983) is implemented in Stata's survey procedures (StataCorp 2005, 264).

Like all methods, linearization has advantages and disadvantages. Some are inherent in the method, whereas others have to do with implementation. Linearization applies to many of the statistics that are calculated from sample surveys like ratios, regression parameter estimates, and specialized combinations that users may construct. The method does not apply directly to estimating the variance of quantiles, like medians, but has been adapted by Francisco and Fuller (1991); however, their method is not yet available in Stata. A package must have a separate linearization estimate programmed for each estimation type (e.g., mean, total, regression parameter). This requirement limits most users to SEs for only the statistics that are preprogrammed.

In principle, one can apply any appropriate variance formula, regardless of how complicated, to the linear approximation for a nonlinear statistic. For example, a two-stage sample with PSUs selected with various probabilities and second-stage units selected by SRS will have a particular variance estimator (for linear estimates) that involves first- and second-order inclusion probabilities of the PSUs (Särndal, Swensson, and Wretman 1992). Stata 9 does allow specifying several sampling stages and some special cases of the variance formulas for such multistage designs.

In variance estimation, formulas for with-replacement sampling of PSUs are often used, even when PSUs were selected without replacement.[2] In with-replacement sampling, one can use a simple variance formula that involves only weighted PSU totals. Using this formula, you only have to specify the PSU in Stata and not any later stages of sampling (for the survey procedures, see Stata's frequently asked questions [FAQs] at http://www.stata.com/support/faqs/stat/#survey). Using with-replacement variance formulas for a without-replacement design typically leads to some degree of overestimation. To counteract this effect, an ad hoc finite population correction factor (fpc) is sometimes inserted. Stata 9 allows fpc specification at each sampling stage. Strictly speaking, these fpc's are appropriate only for stages in which units are selected by SRSs without replacement, but they will help reduce the degree of overestimation.

Variables defining strata and PSUs must be included in a data file so that users can properly calculate linearization variance estimates. To protect confidentiality, the actual strata and PSU identifiers may be masked or otherwise disguised. This method has been used in the NHANES for the 1999–2000 and 2001–2002 public-use datasets (http://www.cdc.gov/nchs/nhanes.htm) and for the National Health Interview Surveys (NHIS) 2003 data file used in section 3.4. Another way to protect confidentiality is the use of replicate weights instead of revealing strata and PSU information. A different method of variance estimation is then necessary, which we will describe in the next section.

4.2 Replicate methods

For a *replication variance estimator*, the sample is broken into subsamples. The desired estimate is computed for each subsample, and the variance is calculated among the subsample estimates. How the subsamples are formed depends on the type of replication variance and can be overlapping or disjointed.

Random groups, jackknife, balanced repeated replication (BRR), and the bootstrap are the replication methods used in survey sampling.[3] Rust and Rao (1996) and Shao (1996) give good reviews of the theory and application of replication in complex surveys. Like linearization, replication applies to linear estimates and nonlinear combinations of linear estimates. Some types of replication can also be applied directly to cases where linearization is difficult, such as estimating the variance of a quantile.

2. This is the case partly because specifying several stages was or is not possible in many packages and partly because the necessary design information (e.g., joint selection probabilities of different units at each design stage) are often not delivered with the survey data for secondary analysis.

3. BRR and jackknife are the methods most often used in practice and are the ones Stata implements.

For reasons noted below, the preferred method of implementation is for a database constructor to calculate the weights and append a series of replicate weights to each record in the file. The user then specifies the method of replication (e.g., jackknife, BRR) and the names of the fields containing the replicate weights to the software package. This method of creating weights can be helpful in protecting confidentiality since strata and PSU codes do not necessarily have to be included in the data file. The database constructor can also repeat the nonresponse, poststratification, or other adjustment steps for the weights separately for every replicate. Generally, adjusting each replicate separately is needed to produce consistent variance estimates when multiple weighting steps are used (Yung and Rao 1996; Valliant 1993). Each weight adjustment, e.g., nonresponse and poststratification, affects the variances of the estimates. Repeating the weight adjustments for each replicate will properly capture their impact on the variance. This repetition is an advantage over most implementations of linearization that tend to account for at most one type of weight adjustment, like poststratification.

Less desirable (usually) is to have a user create replicate weights. This approach requires knowledge of strata and PSU identifiers, and a user generally will not have the detailed information needed to repeat all steps in weight calculation separately for each replicate. In Stata, for example, one option is to let the package divide the sample into jackknife replicates and use the resulting subsets for computing variances. Only basic replicate weight adjustments are made with this approach that do not include nonresponse adjustments or other use of auxiliary data. Consequently, neither the increase in variance due to nonresponse adjustment nor any benefit from poststratification or other use of auxiliary data in weight construction is reflected.

Another advantage of replication is that no analytic derivation is needed to get a linear approximation to a statistic. The software can simply repeat the calculation of a statistic for each replicate and then combine them using the variance formula appropriate to the replication method. This procedure can be used with any Stata estimation command that allows survey weights. For example, a logistic regression of insurance coverage (`NOTCOV`) on age and race (`RACERPI2`) performed using the NHIS data can be run with `svy jackknife: logistic NOTCOV AGE RACERPI2`. This technique is especially handy for user-written ado-files and makes replication more easily applicable than linearization for sophisticated users.

A disadvantage of replication is that it can be computationally intensive. For simple estimates, like means or totals, this is a minor issue, but for iterative procedures like logistic regression, calculation time can be inconveniently long. Another disadvantage is that file sizes can be large if there are hundreds of replicate weights appended to each record. Calculation time and file size, of course, are becoming less of an issue with continual advances in hardware but are still a concern to many analysts. To reduce computation time and file size, database constructors often combine strata and/or PSUs to limit the number of replicates (Rust and Kalton 1987). This method results in fewer degrees of freedom for variance estimates. Database constructors will usually try to minimize the loss of degrees of freedom that this entails.

5 Applications in Stata

Stata includes more analytic procedures for survey data than most other packages. Procedures are available for simple descriptive statistics like means, totals, and proportions. Stata also offers several modeling procedures, including linear regression, instrumental-variables regression, interval and censored regression, binary logistic and probit models, multinomial logistic and probit (ordered and unordered) models, and several other models.

The use of design information and sampling weights in Stata is straightforward and usually has the following structure: (1) specification of the design information (e.g., weights, strata, cluster) followed by (2) a check of the design specification, (3) the estimation itself using the `svy` prefix, and (4) a report of design and misspecification effects. To demonstrate the application of some `svy` commands, we will use publicly available data files for which a short design description is provided. The first example will use explicit design variables, and the second example will describe the same procedure by using replicate weights.

5.1 Linearization

The NHIS series, conducted by the United States National Center for Health Statistics, obtains information about the frequency and dispersion of illness, disabilities, and chronic impairments, as well as the kinds of health services people receive.[4] An area probability sample design is used in which PSUs are geographic areas. Households are selected within PSUs and data are collected by in-person interviews. Three design variables are included in the documentation for the 2003 person-level file[5]: `PSU` indicates the primary sampling units, `STRATUM` the variance strata, and `WTFA` the final analysis weight. Using the command `svyset`, as shown below, this design information is passed to Stata. With no further specification, the default variance estimation method is *Taylor linearization*.

```
. svyset PSU [pweight=WTFA], strata(STRATUM)
      pweight: WTFA
          VCE: linearized
    Strata 1: STRATUM
        SU 1: PSU
       FPC 1: <zero>
```

These `svy` settings will be saved with the dataset, and subsequent users can immediately apply the `svy` commands.

The command `svydes` lists all strata, the total number of PSUs within each stratum (`#Units`), the number of observations in each stratum (`#Obs`), as well as the average, minimum, and maximum number of observations in each PSU (`#Obs per Unit`).

4. Available at http://webapp.icpsr.umich.edu/cocoon/ICPSR-SERIES/00040.xml.
5. ICPSR 4222 Data Documentation.

In the NHIS design, there are two PSUs in each of the 339 strata; within the first stratum, we find 151 observations in the first PSU (`min`) and 158 in the second (`max`).

```
. svydes
Survey: Describing stage 1 sampling units
      pweight: WTFA
          VCE: linearized
     Strata 1: STRATUM
         SU 1: PSU
        FPC 1: <zero>
```

			#Obs per Unit		
Stratum	#Units	#Obs	min	mean	max
1	2	309	151	154.5	158
2	2	347	150	173.5	197
...					
338	2	81	21	40.5	60
339	2	164	80	82.0	84
339	678	92148	6	135.9	389

After these preparations, estimation using the design information requires only the prefix `svy` in front of the estimation command. Looking at the results of `svy: proportion NOTCOV`, we see that an estimated 14.7% of persons in the United States are not covered by any type of health insurance. The 95% confidence interval, which incorporates the design-specific SE, ranges from 14.2% to 15.1%.

```
. svy: proportion NOTCOV
(running proportion on estimation sample)
Survey: Proportion estimation
Number of strata =     339          Number of obs    =   91132
Number of PSUs   =     678          Population size  = 2.8e+08
                                    Design df        =     339
      _prop_1: NOTCOV = Not covered
```

	Proportion	Linearized Std. Err.	Binomial Wald [95% Conf.	Interval]
NOTCOV				
_prop_1	.1465135	.0023286	.1419332	.1510938
Covered	.8534865	.0023286	.8489062	.8580668

The command `estat effects` provides measures for the design and misspecification effects for the estimated proportion not covered by insurance. Here the correct SE is therefore almost twice as big (deft = 1.98) as it would have been from a hypothetical SRS of the same size as we have here. The last column in the output below indicates a meft of 1.87, which means that the width of the resulting confidence intervals would have been underestimated by a factor of 1.87 had the design features been ignored. The confidence intervals are almost twice as wide when they are computed correctly.

```
. estat effects, deft meft
      _prop_1: NOTCOV = Not covered
```

	Proportion	Linearized Std. Err.	Deft	Meft
NOTCOV				
_prop_1	.1465135	.0023286	1.98483	1.87117
Covered	.8534865	.0023286	1.98483	1.87117

In general, once the data are set as survey data, the prefix `svy` will ensure that Stata executes commands while accounting for the survey settings identified by `svyset`. `svy` does not support all estimation commands, but the list is increasing. You can find the current status of commands that `svy` supports with `help survey`.

5.2 Replication

To introduce and discuss variance estimation with replicate weights, we will use data from the PISA study.[6] PISA has a two-stage sampling design with two PSUs (schools) per stratum. Schools were selected proportional to the number of eligible students. Approximately the same number of students were selected within each sampled school. Stratification variables differ by country; for example, in Greece, region and public versus private school were used. However, variables used for stratification purposes are not (or not entirely) included in the PISA data files, nor are identifiers for the PSUs. Instead, the international publicly available data for PISA 2000 are delivered with replicate weights, which implicitly reflect design features.

Again, `svyset` is used for specifying the design and the sampling weight (`w_fstuwt`). Instead of specifying PSUs and strata for the PISA design, the 80 replication weights `w_fstr1-w_fstr80` are used for BRR estimates of the SEs (`vce(brr)`). The replication weights for PISA 2000 are BRR weights using Fay's method (Judkins 1990). The data provider needs to specify whether a Fay constant is needed and if so, what it is. In the standard application of BRR, one PSU is dropped from each stratum and the weights are recalculated with the remaining half-sample. In Fay's method, all PSUs are retained for each replicate but are weighted differently depending on the replicate. In each half-sample, the weights of units in one of the two PSUs in a stratum are multiplied by a factor k, whereas weights in the other PSU are multiplied by $2 - k$. For the PISA data, the Fay adjustment factor needs to be set to $k = 0.5$ in the `svyset` command.

We also specified the option `mse`; with this specification the variance will be computed by subtracting each BRR estimate from the overall sample estimate and squaring the difference. Without the `mse` option, the variance will be centered around the average of the replicate estimates, leading to a somewhat smaller SE. Either method is acceptable, but the `mse` option is more conservative in the sense of giving a larger SE estimate.

6. You can find the data used for this example, as well as more information about the 2000 study and its successors, at http://www.pisa.oecd.org.

```
. svyset [pweight= w_fstuwt], brrweight(w_fstr1-w_fstr80) vce(brr) fay(.5) mse
```

An `svy` prefix will be set in front of Stata commands to obtain correct point estimates and SEs. In the example below, the prefix is now extended to `svy brr`, indicating that BRR will be used. The prefix `svy brr` allows options, one of which we need to use for our dataset.

```
. svy brr: mean pv1read if cnt=="DNK"
(running mean on estimation sample)
BRR replications (80)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..............................
Survey: Mean estimation                 Number of obs    =      4235
                                        Population size  = 47786.1
                                        Replications     =        80
                                        Design df        =        79
```

	Mean	BRR * Std. Err.	[95% Conf.	Interval]
pv1read	496.5618	2.296707	491.9903	501.1333

In our example, we computed the average reading score `pv1read` for students in Denmark.[7] We used the construction `if cnt=="DNK"` to restrict the cases to students in Denmark. This usage of `if` to compare parts of the sample is possible only because samples were drawn independently in each country. Other comparisons between subgroups, e.g., the comparison of reading scores between boys and girls, should not be done with `if`. The options `subpop()` or `over()` should be used instead (see sec. 5.3). The table below shows the means and their confidence intervals for Denmark and the United States, accounting for the designs in each country. Figure 3 already showed the respective means and confidence intervals for the two countries, illustrating that the means were not significantly different. A test of the mean differences `test [pv1read]USA=[pv1read]DNK` confirms this since the p-value on the adjusted Wald test below is 0.3380.

7. PISA uses a form of multiple imputation in which five scores, called plausible values, are created for each test taken by a student. We use only one of five plausible values in our examples. For a more thorough substantive analysis, all five variables should be used and a combined estimate should be reported (Adams and Wu 2002).

```
. svy brr: mean pv1read , over(USA_DNK)
(running mean on estimation sample)
BRR replications (80)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.................................................    50
..............................
Survey: Mean estimation             Number of obs    =    8081
                                    Population size  = 3.2e+06
                                    Replications     =      80
                                    Design df        =      79

          DNK: USA_DNK = DNK
          USA: USA_DNK = USA
```

Over	Mean	BRR * Std. Err.	[95% Conf.	Interval]
pv1read				
DNK	496.5618	2.296707	491.9903	501.1333
USA	503.7085	6.843339	490.0872	517.3298

```
. test [pv1read]USA=[pv1read]DNK
Adjusted Wald test
 ( 1) - [pv1read]DNK + [pv1read]USA = 0
       F(  1,    79) =    0.93
            Prob > F =    0.3380
```

A naive test of mean differences that uses the weights but ignores all other design information can be done with the `mean` and `test` statements as shown below. Since the F statistic for testing the difference in these means has a p-value of 0.0054, this test would have led to the false impression that there is a significant difference between the reading scores of these two countries.

```
. mean pv1read [pweight=w_fstuwt], over(USA_DNK)
Mean estimation                     Number of obs    =    8081

          DNK: USA_DNK = DNK
          USA: USA_DNK = USA
```

Over	Mean	Std. Err.	[95% Conf.	Interval]
pv1read				
DNK	496.5618	1.629446	493.3677	499.7559
USA	503.7085	1.984206	499.819	507.5981

```
. test [pv1read]USA=[pv1read]DNK
 ( 1) - [pv1read]DNK + [pv1read]USA = 0
       F(  1,  8080) =    7.75
            Prob > F =    0.0054
```

5.3 How to make estimates for subgroups

Analysis of subpopulations (also called domains) requires more considerations when estimating SEs. Unless a domain has a fixed sample size, as would be the case if a stratum was a domain, the randomness in the domain sample size should be reflected in the variance estimate. This goal is accomplished by using the `subpop` command, which codes an analysis variable y to zero when a unit is not in the domain. Units in the domain retain their original y values. The variance of the domain estimate is then computed using the recoded y values. An approach that will usually give incorrect SEs is to create a subset file that contains only the cases in the subpopulation and to analyze that subset separately.

One example is the analysis of reading scores for boys within the PISA study. The `svy` option `subpop()` is used to specify an analysis for cases that have a value of 1 in the dummy variable used within the `subpop` option. In the PISA file, `gender` = 1 denotes boys.

```
. svy brr, subpop(gender): mean pv1read
(running mean on estimation sample)
BRR replications (80)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.................................................    50
..............................
Survey: Mean estimation             Number of obs    =   22101
                                    Population size  = 4.8e+06
                                    Subpop. no. obs  =   10895
                                    Subpop. size     = 2.3e+06
                                    Replications     =      80
                                    Design df        =      79
```

	Mean	BRR * Std. Err.	[95% Conf.	Interval]
pv1read	485.5783	5.33818	474.953	496.2037

One potentially confusing phenomenon that can occur when analyzing a subpopulation is that an entire PSU may contain no cases in that domain. This scenario can cause what is known as the singleton PSU problem; i.e., a stratum has only one PSU with observations in the group being analyzed. This situation is most likely to occur in a design where two PSUs are selected per stratum, which is the design for which BRR is often used. When a subpopulation is absent from the sample in a PSU, use of the `subpop` option will cause Stata to code all observations in that PSU to zero for variance calculations, which leads to appropriate estimates of SEs.

Singleton PSUs can also occur for other reasons not related to subpopulation analysis. For example, a data field may have missing values for every observation in a PSU. This situation can happen when a filter question results in a question being asked of few units in a PSU and each unit answers with "Don't know", "Refused", or the like. The Stata FAQs address this situation and others where singleton PSUs are an issue.

6 Summary

This article reviewed key issues in the analysis of complex survey data. Taking survey weights into account as well as information on stratification and clustering is important. Omitting them runs the risk of biased point estimates and erroneous SEs. Weights are necessary when estimating population totals to expand the sample data to the size of the full population. However, estimates of means or proportions can also be biased if weights are not used, as demonstrated in our example that estimated the proportion of hypertensives from NHANES data. When data were collected using a clustered design, SEs are usually larger than those from an unclustered sample. Ignoring clustering in such a case can lead to estimated SEs that are too small, confidence intervals that are too narrow, and hypothesis tests with inflated type I error rates.

The actual implementation of the survey procedures in Stata is straightforward, so thanks to StataCorp, getting the right estimates is easier then ever. Stata, SUDAAN, and R are currently the only major statistical software packages that allow the flexible use of either replicate weights or Taylor linearization to estimate SEs from survey data. Users of survey data should be familiar with both since either may be needed when analyzing publicly available data files.

7 References

Adams, R., and M. Wu. 2002. PISA 2000 technical report. Technical report, OECD, Paris. http://www.pisa.oecd.org/dataoecd/53/19/33688233.pdf.

Binder, D. A. 1983. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 51: 279–292.

Francisco, C., and W. Fuller. 1991. Quantile estimation with a complex survey design. *Annals of Statistics* 19: 454–469.

Groves, R. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2004. *Survey Methodology*. Hoboken, NJ: Wiley.

Judkins, D. R. 1990. Fay's method for variance estimation. *Journal of Official Statistics* 6: 223–239.

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

Kostanich, D., and C. Dippo. 2002. Current population survey: Design and methodology. Technical Report 63RV, Department of Commerce, Washington, DC.

Krewski, D., and J. Rao. 1981. Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics* 9: 1010–1019.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley.

Rust, K., and G. Kalton. 1987. Strategies for collapsing strata for variance estimation. *Journal of Official Statistics* 3: 69–81.

Rust, K., and J. Rao. 1996. Variance estimation for complex surveys using replication. *Statistical Methods in Medical Research* 5: 283–310.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Wiley.

Schnell, R., and F. Kreuter. 2000. Das DEFECT-Projekt: Sampling-Errors und Nonsampling-Errors in Komplexen Bevölkerungsstichproben. *ZUMA-Nachrichten* 47: 89–101.

———. 2005. Separating interviewer and sampling-point effects. *Journal of Official Statistics* 21: 1–23.

Shao, J. 1996. Resampling methods in sample surveys (with discussion). *Statistics* 27: 203–254.

StataCorp. 2005. *Stata 9 Survey Data Reference Manual*. College Station, TX: Stata Press.

Valliant, R. 1993. Post-stratification and conditional variance estimation. *Journal of the American Statistical Association* 88: 89–96.

Yung, W., and J. Rao. 1996. Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology* 22: 23–31.

About the authors

Frauke Kreuter is assistant professor in the Joint Program in Survey Methodology, University of Maryland. Her research interests include interviewer effects on measurement error and nonresponse in surveys. Together with Ulrich Kohler, she is author of the textbook *Data Analysis Using Stata*.

Richard Valliant is a research professor in the Institute for Social Research, University of Michigan, and the Joint Program in Survey Methodology, University of Maryland. His areas of research and application are sampling theory, audit sampling, and analysis of survey data.