Journal of Traumatic Stress, Vol. 21, No. 5, October 2008, pp. 440–447 (© 2008)

Statistical and Methodological Issues in the Analysis of Complex Sample Survey Data: Practical Guidance for Trauma Researchers

Brady T. West
Center for Statistical Consultation and Research at the University of Michigan–Ann Arbor, Ann Arbor, MI

Standard methods for the analysis of survey data assume that the data arise from a simple random sample of the target population. In practice, analysts of survey data sets collected from nationally representative probability samples often pay little attention to important properties of the survey data. Standard statistical software procedures do not allow analysts to take these properties of survey data into account. A failure to use more specialized procedures designed for survey data analysis can impact both simple descriptive statistics and estimation of parameters in multivariate models. In this article, the author provides trauma researchers with a practical introduction to specialized methods that have been developed for the analysis of complex sample survey data.

Many analysts of data collected in sample surveys are comfortable with using standard procedures in statistical software packages to analyze the data; descriptive statistical procedures are used to compute estimates of parameters such as means and proportions for selected variables in specific subgroups being studied, and regression models might be fitted to estimate the relationships between multiple variables. These standard procedures (e.g., the linear regression procedure in SPSS statistical software program) assume that the data being analyzed are arising from a simple random sample from some population of interest, where all sample respondents were randomly selected without replacement from the population, and all respondents had equal probability of being included in the sample (Cochran, 1977, p. 18). Simple random samples have many nice statistical properties: Most important, observations on a given variable are assumed to be independent and identically distributed. Assumptions of simple random sampling are often applied to convenience samples or snowball samples without formal probability designs (where a known probability of inclusion in the sample can be assigned to all population units of analysis).

Unfortunately, in the real world, simple random samples are often quite expensive and difficult to collect. When probability samples from a population of interest are being designed, lists of individual people in a population of interest (or sampling frames) are difficult to locate or produce. Even if a sampling frame enumerating all people in a population of interest is available, the costs of transportation and administration required to collect data from each randomly selected person can be extremely high. As a result, survey statisticians responsible for designing probability samples look for easier, more cost-efficient ways to collect samples from populations. These sample designs, although more cost-efficient and administratively convenient, introduce complications for analysis of the survey data. These designs are often referred to as *complex sample designs*.

In this article, I will introduce readers to issues surrounding the development and analysis of complex samples, including (a) recognizing complex sample designs, (b) the use of sampling weights in analyses of survey data to ensure that computed estimates of desired population parameters are unbiased and representative of the population of interest, (c) the calculation of estimated standard errors for survey estimates that are robust and correctly reflect sampling variability due to the complex design of the sample from which the survey measures were collected, (d) alternative approaches to making inferences based on survey data, (e) the choice of appropriate statistical software to perform unbiased analyses of complex sample survey data, and (f) how to communicate the results of complex sample survey data analyses correctly. All of these points are introduced through an example of analyzing a real survey data set that includes several trauma-related measures: the National Comorbidity Survey-Replication (NCS-R; Kessler et al., 2004).

The author wishes to thank the organizers of the 2007 Conference on Innovations in Trauma Research Methods (CITRM) for the opportunity to present the material in this paper at the conference, and two anonymous reviewers for detailed and thoughtful comments on earlier drafts.

Correspondence concerning this article should be addressed to: Brady T. West, 3550-C Rackham Building, University of Michigan, Ann Arbor, MI 48109. E-mail: bwest@umich.edu.

 DOI: 10.1002/jts.20356

Complex Sample Designs

When designing complex samples, survey statisticians often stratify (or divide) a population of interest into groups that are similar within and different between (in terms of the features being measured in the survey). This process of stratification, which often relies on all possible geographical divisions of the population for administrative convenience, generally has the favorable statistical property of decreasing the standard errors (or increasing the precision) of statistical estimates computed using the survey data. Survey statisticians then assign codes to each respondent indicating his or her stratum, thus producing a stratum variable in the survey data set (for example, the variable containing these codes in the NCS-R data set analyzed in this article is STR). This stratum variable is used by specialized statistical software procedures to compute the proper (and presumably reduced) variance estimates. Readers should note that the stratum variable will include codes for all possible strata of the population; in other words, strata are not randomly selected from the population of interest, and all strata are included in the design.

In addition to stratifying the population of interest, survey statisticians often turn to the sampling of clusters from within the strata; this adds to the complexity of these sample designs. Clusters are groups (or organizations) containing multiple sample elements (or units of analysis in the population, such as trauma patients, if those are the people from whom the survey data will be collected). Examples of clusters might include schools or hospitals, and clusters are nested within sampling strata. For example, after the Oklahoma City bombing, suppose that a survey researcher wished to collect a sample of trauma-related measures from school children in Oklahoma City. The school districts in the city may have represented the sampling strata (ensuring that all possible districts would contribute to the sample), and schools of children may have represented the sampling clusters within the strata. The researcher could then sample schools from within each stratum, and children from within each sampled school.

Clustered samples are often selected for two primary reasons: (a) sampling frames often do not list individual sample elements (e.g., people), but rather clusters of elements (e.g., schools, hospitals); and (b) cluster sampling dramatically reduces the costs of data collection, in that the sampling of clusters within strata (rather than individuals, who might be widely distributed throughout the stratum) facilitates subsequent data collection from samples of individuals within the clusters. There are some statistical trade-offs to cluster sampling, however. Most important, individuals from the same cluster will tend to be similar in terms of the survey measurements of interest, violating the assumption of independent observations for simple random samples. This dependence within a cluster is sometimes referred to as an *intraclass correlation*. This intraclass correlation complicates procedures for the estimation of population parameters and the calculation of standard errors for the estimates, and generally increases the standard errors of the estimates (in contrast to stratification).

In some survey research projects, there are multiple stages of cluster sampling leading to the final selection of a sample element. For example, a complex sample design may feature the random selection of schools from within a stratum (such as a school district), followed by the random selection of classrooms within a school, and followed by the selection of children within the classrooms. These multiple stages of selection also have an impact on the standard errors of survey estimates, in that the standard errors tend to be increased (meaning precision is decreased) due to the multistage cluster sampling. As a result, survey statisticians also assign codes to individuals identifying the sampling clusters to which they belong, for variance estimation purposes; this leads to the inclusion of a cluster variable in public-use survey data sets, which analysts need to identify in preparation for an analysis (the variable containing cluster codes in the NCS-R data set is "SECU"). In multistage cluster sample designs, variance in sample statistics between the primary sampling units (or the first-stage clusters that include multiple lower stage clusters) is of most importance to research analysts, because this variance includes all variance in estimates due to the lower stages of cluster sampling. This is why survey statisticians generally only include a single cluster variable in a survey data set, identifying these first-stage cluster codes.

When analyzing data collected from samples with complex designs, it is the responsibility of the data analyst to understand whether a complex design was used, and identify the important variables containing stratum and cluster codes in the survey data set. Analysts are not responsible for designing these complex samples; they are responsible for analyzing the data correctly given the complex sample design that was used, so that variances of survey statistics are computed properly. This makes it essential for analysts to understand whether or not a survey data set contains these design codes, and use the design codes properly when performing statistical analyses. A thorough review of the documentation accompanying a survey data set, with special attention to any sections concerning estimation of variances or estimation of sampling errors, can be quite helpful with this process.

Complex sample designs tend to result in (a) possible unequal probability of selection into the sample for individual units of analysis, due to the varying sizes of clusters and strata, which is unlike simple random sampling; (b) lack of independence of individual units within randomly sampled clusters, which is also unlike simple random sampling; and (c) increased precision of estimates due to stratification, but decreased precision of estimates due to the use of clustering (in general). These general properties of a complex design led to the development of a unit-less index known as the *design effect* (Kish, 1965), which quantifies the net loss in precision of a survey estimate derived from a complex sample relative to a simple random sample of the same size. The design effect is calculated as follows:

$$\textit{Design Effect} = \frac{Var(estimate)_{complex}}{Var(estimate)_{SRS}} \approx 1 + (b-1) \times ICC$$

The general form of the design effect is a ratio of the variance of a survey estimate according to the complex design (where the variance is estimated using specialized methods that incorporate stratification and clustering) to the variance of a survey estimate had a simple random sample of the same size been collected. In the cases of means and proportions, b corresponds to the average size of the clusters in the population, while *ICC* corresponds to the intraclass correlation of observations within a cluster. Therefore, a higher within-cluster correlation will tend to result in a larger design effect, or increased variance of the survey estimates relative to simple random sampling. Because of nonzero intraclass correlations within clusters, design effects generally tend to be larger than 1 when cluster sampling defines an aspect of the complex sample design. So there are trade-offs: Complex samples are cost-efficient and administratively convenient, but the precision of survey estimates tends to decrease relative to simple random samples.

We now turn to an example of a complex sample survey data set to further introduce these concepts. The working example for this article considers sample data from the NCS-R, which were collected between the years of 2001 and 2003. These data were collected from a nationally representative sample of English-speaking household residents aged 18 years and older in the coterminous United States. The primary aim of the NCS-R was to build on the original National Comorbidity Survey (Kessler, 1990) by continuing to address psychiatric disorders among adults in the U.S. population with improved methods of measurement. The NCS-R sample design was complex in that it involved stratification of the population and multiple stages of cluster sampling leading to the selection of a sample element (a U.S. adult).

Careful assessment of the technical documentation accompanying large survey data sets like the NCS-R is essential for analysts of survey data. The following passage is taken directly from the NCS-R technical documentation (Collaborative Psychiatric Epidemiology Surveys [CPES] online documentation; Heeringa & Berglund, 2008):

> Regardless of whether the linearization method or a resampling approach is used, estimation of variances for complex sample survey estimates requires the specification of a sampling error computation model. CPES data analysts who are interested in performing sampling error computations should be aware that the estimation programs identified in the preceding section assume a specific sampling error computation model and will require special sampling error codes. Individual records in the analysis data set must be assigned sampling error codes that identify to the programs the complex structure of the sample (stratification, clustering) and are compatible with the computation algorithms of the various programs. To facilitate the computation of sampling error for statistics based on CPES data, design-specific sampling error codes will be routinely included in all versions of the data set. Although minor recoding may be required to conform to the input requirements of the individual programs, the sampling error codes that are provided should enable analysts to conduct either Taylor Series or Replicated estimation of sampling errors for survey statistics. (Using CPES, Weighting, Section VII.D)

This passage stresses the need for analysts to use specialized statistical software procedures when analyzing the NCS-R data. Readers should note the comments about the complex design effects and the need to take the design structure of the NCS-R sample into account when using specialized software to perform analyses of the survey data. We will revisit the NCS-R example in all of the subsequent sections of this article.

Sampling Weights

Complex sample designs often result in unequal probabilities of selection into the sample for population units of analysis (e.g., people). If certain individuals have a higher probability of being included in a sample than others, population estimates based on that sample design may be biased. In addition, sampled individuals may choose not to respond to a survey, even if incentives are used as a part of the research, and the distribution of a sample may not be in alignment with known population characteristics (e.g., a sample might be 30% men and 70% women, when the population of interest is known to be 50% men and 50% women). These three complications result in the need to use sampling weights when analyzing the survey data, so that these potential sources of bias in the survey estimates are reduced (or even eliminated). Here a very general description of sampling weights is provided; for additional technical information on the computation of sampling weights, interested readers can refer to Kish (1965), Cochran (1977), or Lohr (1999).

Computed sampling weights for people who respond to a survey incorporate (a) unequal probability of selection (people with a lower probability of being included get more weight when estimates are computed); (b) nonresponse rates (people belonging to groups that are not as likely to respond to the survey get more weight when estimates are computed); and (c) poststratification factors, or adjustments to match population distributions across certain strata (e.g., men and women). The first two components of a sampling weight are generally computed by taking the inverse of an individual's probability of selection and the inverse of the response rate for a certain group to which an individual belongs, and multiplying the two components together. For example, consider a survey respondent who has a 1/100 chance of being included in a sample, and belongs to a demographic group where 50% of people who were sampled actually responded. This individual would have a sampling weight of $100 \times 2 = 200$, meaning that they represent themselves and 199 other people. This weight might then be further adjusted by a poststratification factor to match known population distributions.

The final computed sampling weight is also generally included in a survey data set as a weight variable, in addition to the stratum and cluster variables (the appropriate weight variable in the NCS-R data set for the examples in this article is FINALP2W). The analyst has the responsibility of identifying this final weight variable so that specialized estimation routines in statistical software can incorporate the sampling weights when computing unbiased estimates of population parameters such as means and regression coefficients. A failure to incorporate sampling weights in analyses can result in severely biased estimates of important parameters for a population of interest.

In the NCS-R example, a subsample of 5,692 of the originally sampled 9,282 respondents (those indicating a lifetime mental health disorder in addition to a sample of other respondents screening as completely healthy) answered a second set of questions in the NCS-R (Part II) measuring risk factors and additional disorders, including posttraumatic stress disorder. The original NCS-R sampling weights were adjusted for this additional subsampling, and the NCS-R technical documentation explains to analysts that the FINALP2W variable contains these adjusted weights that should be used for all analyses involving Part II respondents.

Variance Estimation

Survey statisticians compute sampling weights for the information provided by sample respondents to enable survey analysts to compute unbiased estimates of population parameters. What this means is that on average (or in expectation), the estimates (or statistics) will equal the population parameter of interest. This makes the estimates unbiased in theory; in practice, different samples from a population will result in different estimates of a population parameter that "bounce around" the true value of the parameter. Generally speaking, the standard error of a survey estimate describes how far away an estimate from any one sample tends to be from the expected value of an estimate (the parameter it is estimating), on average. Methods for variance estimation in analyses of complex sample survey data are used to compute estimates of the standard errors of complex sample survey statistics.

The need to compute estimates of standard errors often raises questions among researchers. Why not just compute the true value of a variance (and a standard error, which is the square root of the variance) for a given sample estimate of a population parameter? Unfortunately, when analyzing data collected from a sample with a complex design, true values of population parameters are needed to compute the true theoretical variances of sample statistics, and we will never know these "true" values. As a result, good approximations of the theoretical standard errors based on sample estimates of the required population parameters are used to estimate the standard errors in a robust way that accounts for the stratification, clustering, and sample weighting underlying a given complex design. Survey statisticians research variance estimators that are as close to being unbiased estimators of the true theoretical variances for sample statistics as possible.

In general, the estimated variance of a given survey statistic is based on within-stratum variances in sample statistics (such as totals) between first-stage sampling clusters. As a result, at least two clusters are required per stratum for variance estimates to be calculated; otherwise, within-stratum variances (and therefore overall variance estimators) cannot be calculated. If only a single cluster is found in a particular stratum by the software processing the sample design variables, errors or warning messages may appear that could prevent the software from reporting estimated standard errors for survey statistics of interest. Survey sampling statisticians therefore do their best to provide users of survey data sets with variables identifying the strata and clusters to which each respondent belongs such that each stratum code has at least two associated cluster codes for variance estimation purposes.

Three of the mathematical methods most frequently used to estimate the variances of complex sample survey statistics include Taylor series linearization, jackknife repeated replication, and balanced repeated replication. Taylor series linearization, which is the default variance estimation technique in many statistical software procedures designed for survey data analysis, first "linearizes" complex, nonlinear statistics (such as ratios of sample totals) into linear functions of sample totals, using results from calculus (Taylor series approximations). Variances of these linearized approximations of the original nonlinear statistics can then be calculated by using simpler known formulas for the variances and covariances of sample totals within strata. Each different survey statistic will have a unique variance estimator based on Taylor series linearization. In contrast, jackknife repeated replication and balanced repeated replication are very general techniques for variance estimation based on resampling, where the same methods are used to estimate variances regardless of the form of the survey statistic. Essentially, replication methods involve drawing repeated samples from the available survey data set, calculating the statistic of interest for each of the repeated samples, and then describing the variance of the repeated statistics around the overall statistic based on the overall sample. Software packages with procedures available for complex sample-survey data analysis are at various stages of incorporating these three techniques; however, Taylor series linearization is generally sufficient for most practical applications.

These variance estimation methods will tend to result in robust overestimates of what the true theoretical variances should be (keeping in mind that stratification will generally decrease variances while clustering and weighting will increase variances), and this is a better scenario in practice than underestimating the variances, or overstating the precision of the survey estimates. As a result, the methods tend to provide conservative estimates of variances for the survey statistics, protecting the survey analyst against the risk of overstating the precision of survey estimates. The choice of the variance estimation method generally depends on software availability; Taylor series linearization is currently more widely

available than the replication techniques. Linearization and jackknife repeated replication are generally more accurate and more stable when the survey statistics represent functions of means, whereas balanced repeated replication has been shown to be better for the estimation of quartiles. As mentioned earlier, Taylor series linearization is generally a reasonable choice for most applications. For more details, interested readers can refer to a simulation study by Kovar, Rao, and Wu (1988) that compares the variance estimation techniques.

Analysts have the important responsibility of identifying the appropriate complex design variables in a survey data set (e.g., FINALP2W, STR, and SECU in the NCS-R data set) and using them properly in statistical software procedures designed for these analyses, so that robust estimates of standard errors incorporating the complex design features will be calculated correctly.

Design-Based Versus Model-Based Inference

Generally speaking, statistical inference refers to the art of making statements about a population based on a sample of data collected from that population. Design-based inference refers to an approach to making inferences using survey data that utilizes confidence intervals determined by estimates of survey statistics and their standard errors. The estimates and standard errors are determined using nonparametric methods (without assuming probability distributions for the variables) that incorporate features of the complex sample design and rely on how representative the sample is of the population of interest. Researchers examine the computed confidence intervals (CIs) to determine whether they include the hypothesized value of a population parameter of interest. One would interpret a 95% CI (corresponding to a type I error rate of .05) for a population parameter by stating that 95% of CIs constructed in the exact same way across repeated samples would include the true population parameter.

In contrast, model-based inference relies on assumptions of specific probability distributions for the variables being analyzed. This method of inference is extremely powerful if the assumed probability distributions for the variables are correct. One example of a model-based approach to the analysis of survey data is multilevel (or mixed-effects) modeling, where the effects of clustering are treated as random effects in linear models (e.g., Pfefferman, Skinner, Holmes, Goldstein, & Rasbash, 1998). This article's primary focus is on methods for design-based inference, or robust, nonparametric inference that directly recognizes the complex design features of a representative probability sample. For further details on these two forms of inference for survey data, interested readers can refer to Hansen, Madow, and Tepping (1983).

Current Software Options

A wide variety of software procedures capable of analyzing survey data are currently available to survey researchers. Considering commercial statistical software packages first, the SAS software (Version 9.1.3; SAS Institute, Cary, NC) currently offers procedures for descriptive analysis of continuous and categorical variables (PROC SURVEYMEANS and PROC SURVEYFREQ), in addition to two procedures for regression modeling (PROC SURVEYREG and PROC SURVEYLOGISTIC). The SPSS software (Version 16; SPSS Inc., Chicago, IL) currently offers the complex samples add-on module, which needs to be purchased in addition to the base SPSS package; this module enables several descriptive analyses, in addition to regression modeling for continuous, categorical, count, and time-to-event outcomes. Here we will consider the complex samples module for the NCS-R analysis. The Stata software package (Version 10; StataCorp LP, College Station, TX) and the SUDAAN software package (Version 9.0.1; Research Triangle Institute, Research Triangle Park, NC) currently offer the widest variety of procedures for survey data analysis, including survival analysis and estimation of percentiles. Other software packages including WesVar (http://www.westat.com/wesvar/) and IVEware (http://www.isr.umich.edu/src/smp/ive/) are freely available for survey data analysis, and the IVEware package is especially useful for missing data problems (which are beyond the scope of this article).

Returning to the NCS-R analysis example, we define two research objectives: (a) to estimate the percentage of the adult U.S. population that has ever participated in combat; and (b) to estimate the percentage of older (age ≥ 50) men in the United States that has ever participated in combat. The second analysis objective defines a subclass analysis, and more details on these types of analyses will be provided in the next section. We consider the SPSS software for this example, although this example analysis could be performed using any of the aforementioned statistical software packages. The following NCS-R variables will be considered in the analysis (in addition to the FINALP2W, STR, and SECU variables, representing the sampling weights and the stratum and cluster codes, respectively): PT1 (Has the respondent ever participated in combat, as a member of a military or organized nonmilitary group?), SC1 (respondent's age), and SC1_1 (respondent's gender).

The first step for SPSS users looking to analyze a complex sample-survey data set is to purchase the complex samples add-on module: readers can visit http://www.spss.com/complex_samples for more information. Other general-purpose statistical software packages like SAS, Stata, and SUDAAN have procedures for the analysis of complex sample survey data integrated into their base statistical packages. With the NCS-R data set open in SPSS, the next step is to define the complex design features of the NCS-R. From the SPSS menus, one can select Analyze → Complex Samples → Prepare for Analysis. . . This allows users to define what is known in SPSS as a plan file, which has an extension of .csaplan. Users first choose the option to create a plan file, and then enter a name for the file before clicking Next to proceed. At this point, users select the variables containing information on the strata, clusters, and sampling weights that are critical for a design-based

PT1: Evr in combat military/non-military

		Estimate	Standard Error	95% Confidence Interval		Design Effect	Unweighted Count
				Lower	Upper		
Population Size	1	277.330	23.122	230.667	323.993	2.026	243
	5	5413.310	247.467	4913.902	5912.718	232.024	5446
	Total	5690.640	251.750	5182.589	6198.691	.	5689
% of Total	1	4.9%	.4%	4.1%	5.8%	2.102	243
	5	95.1%	.4%	94.2%	95.9%	2.102	5446
	Total	100.0%	.0%	100.0%	100.0%	.	5689

Figure 1. SPSS output from the first combat example, showing that an estimated 4.9% of U.S. adults have ever participated in combat before (variable PT1 = 1).

analysis of the survey data, and click Next. The standard sampling error calculation model selected in the next screen is WR (or with replacement), which assumes that the first-stage clusters (not the individual sample elements) have been selected with replacement from within the first-stage strata (an assumption that primarily facilitates variance estimation; see Cochran, 1977, for more details). Users then click Finish to finish the definition of the sampling plan.

Next, one defines the analysis to be performed. Users can select Analyze → Complex Samples → Frequencies. . . to request estimates of percentages for categorical variables. One first has to identify the sampling plan file to be used in the analysis, which was created in the previous steps. Next, the variable for which weighted percentages are desired (PT1, or the indicator of participation in combat) is selected for frequency tables analysis, and selected statistics (e.g., standard errors, 95% CIs, design effects) can be requested for the output. After these options have been selected, the user can run the analysis procedure by clicking OK (or pasting and running the SPSS syntax). This generates the SPSS output displayed in Figure 1.

The weighted estimate of the percentage of U.S. adults having participated in combat is 4.9% (Linearized $SE = 0.4\%$, 95% CI = 4.1%–5.8%). A null hypothesis that 5% of U.S. adults have participated in combat would not be rejected, but a null hypothesis stating that 10% of adults have participated in combat is not supported by these results. The 95% CI reflects expected sampling variability in the estimate given the complex design features of the NCS-R sample, and the estimated standard error for the estimate is computed using Taylor series linearization. The unweighted estimate of this percentage, ignoring the complex design features and performing a standard frequency table analysis in SPSS (Analyze → Descriptive Statistics → Frequencies), is 4.3%, which indicates that a more standard analysis would have resulted in an estimate that was biased low. Further, the estimated design effect (DEFF) is 2.1, suggesting that the complex design of the NCS-R sample increased the variance of this estimate by around 110% relative to a simple random sample of the same size (indicating a loss in the precision of the estimate due to the complex design). This finding is quite common for design effects, and is mainly driven by the effects of clustering and weighting, as discussed earlier.

Subclass Analyses

Analysts of survey data are often interested in restricting inferences to a specific subpopulation, or subclass, of the population of interest. For example, one may wish to restrict the estimation of a population parameter to older men. In practice, subclass analyses of complex sample survey data are dangerous because it is very easy for analysts to incorrectly (and permanently) delete those cases that do not fall into the subclass from the data set when performing the analysis. The two primary problems with this type of conditional approach to the subclass analysis are (a) the subclass sample size is treated as fixed, when it should actually be treated as a random variable (i.e., the subclass sample size will vary in theory from one sample to another); and (b) deleting cases not falling into the subclass may result in entire first-stage sampling clusters being deleted from the analysis, if the subclass by random chance is not represented in a given cluster in a given sample.

Estimated standard errors of survey statistics should reflect theoretical sample-to-sample variance in the statistics based on the original complex sample design. In other words, across repeated samples, how far off (on average) will estimates be from the true population parameter of interest? If clusters from the original sample design are deleted in a conditional subclass analysis, when in theory the subclass might have appeared in those clusters, standard errors will be underestimated. The statistical software used to analyze the subclasses will not know that the deleted clusters were ever part of the original complex design. In addition, the software will treat the subclass sample size as being fixed from sample to sample when calculating the standard errors, when sample-to-sample variability in the subclass sample sizes should be incorporated into the standard errors. This also leads to underestimation of standard

PT1: Evr in combat military/non-military

oldmale			Estimate	Standard Error	95% Confidence Interval Lower	95% Confidence Interval Upper	Design Effect	Unweighted Count
0	Population Size	1	72.700	13.169	46.123	99.277	2.415	77
		5	4663.090	232.104	4194.684	5131.496	63.951	4853
		Total	4735.790	232.953	4265.672	5205.908	68.260	4930
	% of Total	1	1.5%	.3%	1.1%	2.2%	2.478	77
		5	98.5%	.3%	97.8%	98.9%	2.478	4853
		Total	100.0%	.0%	100.0%	100.0%	.	4930
1	Population Size	1	204.630	20.721	162.813	246.447	2.176	166
		5	750.220	60.064	629.006	871.434	5.537	593
		Total	954.850	66.025	821.606	1088.094	5.483	759
	% of Total	1	21.4%	2.0%	17.6%	25.8%	2.341	166
		5	78.6%	2.0%	74.2%	82.4%	2.341	593
		Total	100.0%	.0%	100.0%	100.0%	.	759

Figure 2. SPSS output from the application of the combat example to selected population subclasses (men with age ≥ 50 are represented by OLDMALE $= 1$), showing that an estimated 21.4% of men 50 years of age or older have ever participated in combat before (as opposed to an estimated 1.5% of other adults).

errors. Conditional subclass analyses can also result in the situation where the software only recognizes a single cluster in a given sampling stratum, and different software packages react differently to this problem because within-stratum variance between the clusters cannot be estimated.

To prevent these critical problems that can arise when performing subclass analyses, survey data analysts should perform "unconditional" subclass analyses. Specialized options in statistical software procedures allow users to define an indicator variable in the survey data set for the subclass of interest, equal to 1 for cases in the subclass, and 0 otherwise. By processing this indicator variable, the software can recognize the full complex design of the sample, treat the subclass sample size as random, and proceed with the estimation of the statistics and standard errors according to the full complex design (clusters containing no subclass elements are still recognized in the variance calculations); interested readers can refer to West, Berglund, and Heeringa (in press) or Cochran (1977) for more details. We consider this unconditional approach by analyzing older (age ≥ 50) men in the NCS-R data set, and estimating what proportion of this subclass has even participated in combat.

First, we compute an indicator variable (OLDMALE), equal to 1 for respondents in this subclass, and 0 otherwise. Then, we select Analyze → Complex Samples → Frequencies. . . to request estimates of the proportion for this specific subclass. We request frequency tables for PT1 (the indicator of participation in combat), and then move the indicator variable for older men into the subpopulations window. This will request an unconditional subclass analysis for the older men. Running this analysis produces the SPSS output shown in Figure 2.

Note in Figure 2 that an estimated 21.4% of men 50 years of age or older (Linearized $SE = 2.0\%$, 95% CI $= 17.6\%$–25.8%) ever participated in combat, which is quite different from the estimate of 4.9% for the overall population. There is a larger margin of error for this smaller subclass, which is quite common in subclass analyses; the Linearization-based standard error of the estimate in this case is 2.0%, compared to 0.4% for the overall analysis. The unweighted estimate for this subclass is 21.8%, suggesting that there is not a great deal of bias in the unweighted estimate. Further, the design effect for this estimate is 2.3, suggesting a loss of precision relative to a simple random sample of the same size from this subclass. Given statistics of primary interest, survey statisticians work to minimize this design effect when designing complex samples.

Analysts may be interested in statistically comparing selected subclasses in terms of important descriptive parameters (e.g., means or proportions) when analyzing a survey data set. For example, one may wish to compare respondents with combat experience with respondents not having combat experience in terms of the prevalence of PTSD. In this case, a design-based version of the chi-square test could be performed, comparing these two groups in terms of the prevalence of PTSD. The design-based analog of the χ^2 test is the Rao-Scott χ^2 test (Rao & Scott, 1984), which has been widely implemented in the different software packages mentioned in this article. In this case, no subclass indicator variable would be required because the two groups of interest would

define the entire population being studied. If one were to adjust for covariates when comparing the groups in terms of this prevalence, a logistic regression model should be fitted, and specialized procedures capable of fitting logistic regression models to survey data are also widely implemented in the different statistical software packages. Regardless of the model to be fitted or the subclass comparison of interest, the important point for analysts is that a specialized procedure be used to perform the analysis that can correctly incorporate complex sample design features when generating estimates, standard errors, and test statistics used for making inferences about populations of research interest.

Conclusion

A broad and general introduction to the analysis of complex sample survey data has been presented in this article. Analysts of survey data have several issues to keep in mind before beginning the analysis of a complex sample survey data set:

1. What was the nature of the complex sample design, and are there codes for strata and clusters included in the survey data set?
2. What is the correct sampling weight variable to be used in the analysis for calculating unbiased estimates of population parameters?
3. Is a software procedure being used that can accommodate the complex sample design features in the analysis?
4. What method of variance estimation is the software procedure using to calculate design-based estimates of standard errors (e.g., Taylor series linearization)?
5. Is the analysis of interest focused on a subclass of the population from which the complex sample survey data were collected, and has an indicator variable been computed for this subclass enabling unconditional subclass analyses?

Example analyses in this article were presented using the complex samples module of the SPSS statistical software (Version 16), but the analyses can be easily repeated using specialized procedures for survey data analysis in several other statistical software packages (e.g., SAS, Stata, and SUDAAN). The author can be contacted for examples of code that can be used to perform the analyses in these other software packages.

The scientific design of cost-efficient probability samples and the collection of survey data from the resulting samples in survey research projects like the NCS-R cost granting agencies (and tax payers) millions of dollars each year. A failure of data analysts to use appropriate statistical software procedures for unbiased analyses of the survey data essentially negates the substantial money and effort used to collect the data, and can result in scientific publications that present skewed pictures of populations of research interest. This presentation has attempted to clarify these issues for analysts of complex sample survey data in a relatively heuristic and practical manner. Several additional references can be consulted for additional technical details on these analysis approaches, including a practical handbook by Lee and Forthofer (2006), and theoretical articles on variance estimation by Rust (1985) and Binder (1983).

REFERENCES

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

Cochran, W. G. (1977). Sampling techniques (3rd ed.). New York: Wiley.

Hansen, M., Madow, W., & Tepping, B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. Journal of the American Statistical Association, 78, 776–793.

Heeringa, S. G., & Berglund, P. (2008). National Institutes of Mental Health (NIMH) Collaborative Psychiatric Epidemiology Survey Program (CPES) data set. Integrated weights and sampling error codes for design-based analysis. Retrieved June 5, 2008, from http://www.icpsr.umich.edu/cocoon/cpes/using.xml?section=Weighting#VII.+Procedures+for+Sampling+Error+Estimation+in+Design-based+Analysis+of+the+CPES+Data

Kessler, R. C. (1990). National Comorbidity Survey: Baseline (NCS-1), 1990–1992 [Data file]. Conducted by University of Michigan, survey research center (ICPSR06693-v4). Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Kessler, R. C., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S., Hiripi, E., et al. (2004). The US National Comorbidity Survey Replication (NCS-R): Design and field procedures. International Journal of Methods in Psychiatric Research, 13, 69–92.

Kish, L. (1965). Survey sampling. New York: Wiley.

Kovar, J. G., Rao, J. N. K., & Wu, C. F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. Canadian Journal of Statistics, 16, 25–45.

Lohr, S. L. (1999). Sampling: Design and analysis. Pacific Grove, CA: Duxbury Press.

Lee, E. S., & Forthofer, R. N. (2006). Analyzing complex survey data (2nd ed.). Quantitative Applications in the Social Sciences, 71. Newbury Park, CA: Sage.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society, 60, 23–40.

Rao, J. N. K., & Scott, A. J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. Annals of Statistics, 12, 46–60.

Rust, K. (1985). Variance estimation for complex estimation in sample surveys. Journal of Official Statistics, 1, 381–397.

West, B. T., Berglund, P., & Heeringa, S. G. (in press). A closer examination of subpopulation analysis of complex sample survey data. The Stata Journal.