The Stata Journal (2008)
8, Number 4, pp. 520–531

A closer examination of subpopulation analysis of complex-sample survey data

Brady T. West
Center for Statistical Consultation and Research
University of Michigan, Ann Arbor
Ann Arbor, MI
bwest@umich.edu

Patricia Berglund
Institute for Social Research
University of Michigan, Ann Arbor
Ann Arbor, MI
pberg@umich.edu

Steven G. Heeringa
Institute for Social Research
University of Michigan, Ann Arbor
Ann Arbor, MI
sheering@umich.edu

Abstract. In recent years, general-purpose statistical software packages have incorporated new procedures that feature several useful options for design-based analysis of complex-sample survey data. A common and frequently desired technique for analysis of survey data in practice is the restriction of estimation to a subpopulation of interest. These subpopulations are often referred to interchangeably in a variety of fields as subclasses, subgroups, and domains. In this article, we consider two approaches that analysts of complex-sample survey data can follow when analyzing subpopulations; we also consider the implications of each approach for estimation and inference. We then present examples of both approaches, using selected procedures in Stata to analyze data from the National Hospital Ambulatory Medical Care Survey (NHAMCS). We conclude with important considerations for subpopulation analyses and a summary of suggestions for practice.

Keywords: st0153, survey data analysis, statistical software, complex sample designs, subpopulation analysis

1 Introduction

Health care researchers, epidemiologists, and social scientists analyzing survey data from probability samples with complex, multistage designs (where the designs incorporate stratification and clustering of the study population) are often interested in focusing their analyses on specific subgroups of the full population sample (e.g., Midwest region residents, Hispanics, or males and females). The survey literature and the documentation for software that implements design-based analyses of survey data use a variety of terms to refer to these subgroups, including subclasses, subpopulations, domains, and subdomains. In this article, we refer to these subgroups as *subpopulations*. The primary objective of this article is to provide analysts using complex-sample survey datasets with some practical guidance on appropriate approaches to the analysis of subpopulations of survey datasets, specifically when using the Stata software.

The past decade has seen tremendous growth in the availability and power of software for the analysis of survey data. Procedures designed for these specialized analyses have recently incorporated additional options for performing subpopulation analyses of survey datasets. In particular, the Stata software package currently has options for subpopulation analyses implemented in all its survey data analysis procedures (StataCorp 2007). These developments have made it possible for analysts of survey datasets to perform complex analyses quickly and easily, but they have also introduced the opportunity to make critical mistakes when the software options for subpopulation analysis are not used properly. Few articles have provided survey analysts with practical guidance about performing subpopulation analyses; a recent article by Kreuter and Valliant (2007) dedicated a section to the *subgroup* capabilities within the Stata procedures for survey data analysis. In this article, we aim to focus in more detail on the practical issues underlying subpopulation analysis of complex-sample survey data.

Section 2 of this article presents a heuristic discussion of the conceptual differences in "conditional" and "unconditional" approaches (Cochran 1977) to the analysis of subpopulation data in complex-sample survey datasets. Section 3 then presents a motivating example using data from the 2004 National Hospital Ambulatory Medical Care Survey (NHAMCS), introducing the problems with variance estimation that can arise when (incorrectly) following conditional approaches. Section 4 contains a discussion of issues related to the calculation of degrees of freedom when performing subpopulation analyses, referring to results from the motivating example. Finally, section 5 presents a summary discussion of the issues presented.

2 Alternative approaches to subpopulation estimation and inference

When working with simple random samples (SRSs) of survey populations of interest, practicing survey data analysts interested in analyzing subpopulations simply need to restrict the dataset to those observations falling into the subpopulation of interest before performing analyses. Conceptually, an SRS from a population of interest will include SRSs of subpopulations as well. In this situation, users of Stata might use the `if` qualifier to restrict analyses to those cases satisfying a certain condition (e.g., males only), or users might create and save a dataset containing only cases in a subpopulation of interest. In this article, we refer to this analytic approach as a conditional approach: the analysis is "conditioned" on the sample of observations obtained for the subpopulation under the larger SRS sample selection.

We consider "complex" sample designs in this article that, at a minimum, involve stratification of the sample and may also include features such as clustering of sample units and disproportionate probabilities of selection for observational units (i.e., weighting in analysis). The issues discussed in this article are relevant for any survey datasets arising from a sample with a stratified design, where subpopulation sample sizes within the strata are not known at the time the sample is selected. We present example subpopulation analyses of real survey data arising from a sample with a stratified multistage

design (e.g., Cochran [1977]), where first-stage primary sampling units (PSUs) were selected with replacement from within first-stage sampling strata, and sample elements were randomly selected without replacement from within the PSUs at the second stage and subsequent stages. When working with survey datasets collected from samples with complex designs, which at a minimum incorporate stratification of target populations, subpopulation analyses are not as straightforward as in the SRS case. Perhaps surprisingly to some analysts, taking the conditional approach to subpopulation analyses of complex-sample survey data has the potential to result in incorrect standard errors for survey estimates.

One common pitfall when performing conditional subpopulation analyses of survey datasets that arise from samples featuring stratification *and* clustering is the deletion of PSUs that define a part of the complex design. Most subpopulations of survey datasets could hypothetically appear in all strata and all PSUs (or clusters) in any given complex sample. There is the potential, however, that a given complex sample by random chance does not include elements from a subpopulation of interest (e.g., Asian males age 50 and above) in one or more of the PSUs (even though the subpopulation *could theoretically appear* in those PSUs). If an analyst follows a conditional approach in this situation, one or more of the PSUs defining the original complex sample design used to collect the data could be deleted from the dataset, because that PSU does not include any sample elements in the subpopulation. The software used for analysis would have no idea that an original PSU based on the complex design ever existed. As most modern analysts of complex-sample survey datasets know, allowing software to fully acknowledge the strata and PSUs defining a complex sample is essential for correct variance estimation, and following the conditional approach can prevent this from happening. This problem is generally not relevant when working with stratified samples that do not feature the selection of clusters, because subpopulations that are missing from an entire stratum in a sample more than likely do not exist in that stratum.

Mathematically, estimates of standard errors for statistics from complex-sample survey datasets need to take into account sample-to-sample variability in the statistics of interest based on the original complex design. When focusing on subpopulations, methods of variance estimation for all point estimates of survey statistics should be unconditional in that they take the full complex design of a sample into account, and they should not be restricted to only those sample cases falling into the subpopulation (the conditional approach). The major mathematical motivation behind this principle is that, in practice, observed subpopulation sample sizes for strata (and PSUs, when applicable) are *random variables*, and the true subpopulation size is not known (and needs to be estimated). From one sample to another, subpopulation sample sizes within strata (and PSUs, when applicable) will vary in size, and this variance needs to be taken into account when variance estimates are calculated.

To illustrate the importance of following the unconditional approach mathematically, we consider the variance of a sample total; this is because of the fact that all survey statistics (and their variances) are ultimately expressed as functions of sample totals for variance estimation purposes, especially when using the Taylor series linearization technique (e.g., Wolter [1985]). We denote design strata by h ($h = 1, 2, \ldots, H$), first-

stage PSUs within strata by α ($\alpha = 1, 2, \ldots, a_h$), and sample elements within PSUs by i ($i = 1, 2, \ldots, n_{h\alpha}$). Taking into account factors such as unequal probability of selection, nonresponse, and possibly poststratification, the sampling weight for element i is denoted by $w_{h\alpha i}$. We refer to a specific subpopulation by S. An estimate of the total for a variable Y in a subpopulation S is computed as follows (Cochran 1977):

$$\widehat{Y}_S = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}\, I_{S,h\alpha i}\, Y_{h\alpha i}$$

where I represents an indicator variable equal to 1 if sample element i belongs to subpopulation S, and 0 otherwise. The closed-form analytic equation for the variance of this subpopulation total can be written as follows:

$$\begin{aligned} \mathrm{var}(\widehat{Y}_S) = & \sum_{h=1}^{H} \frac{a_h}{(a_h - 1)} \\ & \left\{ \sum_{\alpha=1}^{a_h} \left(\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}\, I_{S,h\alpha i}\, y_{h\alpha i} \right)^2 - \frac{\left(\sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}\, I_{S,h\alpha i}\, y_{h\alpha i} \right)^2}{a_h} \right\} \end{aligned} \tag{1}$$

Equation (1) shows that the variance of the subpopulation total is calculated by summing the between-cluster variance in the subpopulation totals within strata, across the H sample strata. The equation also shows how the indicator variable is used to ensure that all sample elements (and their design strata and PSUs) are recognized in the variance calculation; this emphasizes the need for the software to recognize all the original design strata and PSUs. In this way, sample-to-sample variability in the estimation of the total due to the subpopulation sample size being a random variable is captured in the variance calculation. Also if all $n_{h\alpha}$ elements within a given stratum denoted by h and PSU denoted by α do not belong to the subpopulation S (although elements from that subpopulation theoretically could belong to that PSU in any given sample), the PSU will still contribute to the variance estimation: the PSU helps to define the total number of PSUs within stratum h (a_h).

For more information on variance estimation approaches used by the Stata software for commonly estimated survey statistics, we refer readers to Eltinge and Sribney (1996a, 1996b, 1996c, 1996d, 1996e). Proper analysis methods for subpopulations of survey data have been well established in the survey methodology literature (Rao 2003; Korn and Graubard 1999; Lohr 1999; Fuller et al. 1989; Cochran 1977; and Kish 1965), and interested readers can consult these references for more general information on estimation of survey statistics and related variance estimation techniques. Choosing to perform either a conditional or unconditional subpopulation analysis will only impact variance estimates; weighted estimates of survey statistics will not be affected, and this will be illustrated in section 3.

3 A motivating example: Analyses of data from the 2004 NHAMCS Emergency Department Sample

We now consider a motivating example in Stata to illustrate the importance of performing unconditional subpopulation analyses of complex-sample survey data. We consider data from the 2004 NHAMCS. Briefly, the NHAMCS collects data from annual, cross-sectional, four-stage probability samples of visits to randomly selected, noninstitutional, general and short-stay U.S. hospitals with emergency departments or outpatient departments (McCaig and McLemore 1994). Dizziness is one of the most common reasons that patients present to physicians' offices, hospital outpatient departments (OPDs), and emergency departments (EDs) in the United States (Burt and Schappert 2004). In this example, we consider a very specific subpopulation: visits to EDs (only) by elderly (age $\geq$ 60) African-American males. This subpopulation is somewhat specific but can still theoretically appear across the strata and PSUs in a nationally representative sample design involving stratification and clustering. We seek to estimate the percentage of these visits with dizziness or vertigo as a reason for the visit.

NHAMCS analytic guidelines (e.g., Hing et al. [2003]) specifically indicate that analyses focusing only on ED visits require analysts to combine separate datasets for the ED and OPD visits in a given year and perform unconditional subpopulation analyses, treating ED visits as representing a subpopulation of the full NHAMCS sample (which includes OPD visits). We combined the 2004 ED and OPD datasets, publicly available from the National Center for Health Statistics, and computed an indicator variable in Stata for our specific subpopulation of interest in the combined dataset:

```
. generate subc = (settype == 3 & sex == 2 & agecat == 5 & race == 2)
```

Here `settype` = 3 represents ED visits, `sex` = 2 represents males, `agecat` = 5 represents ages greater than or equal to 60, and `race` = 2 represents African Americans. We then used commands in Stata to examine the complex design of the 2004 NHAMCS sample, referring to the variables containing stratum (`cstratm`) and ultimate cluster (`cpsum`) codes computed by National Center for Health Statistics staff for variance estimation purposes:

```
. svyset cpsum [pweight = patwt], strata(cstratm)
. svydescribe
```

The resulting output indicated 8 strata, containing between 6 and 86 PSUs each. All together, there were 8 strata and 294 PSUs defining the full complex design of the 2004 NHAMCS.

We then computed an indicator variable for visits where any one of the reasons for the visit was vertigo and/or dizziness (`dizzyrfv`). First, we considered a conditional approach to estimating the percentage of visits in this subpopulation for dizziness and/or vertigo in 2004, restricting the dataset to the visits of interest (the `se`, `ci`, and `percent` options request that Stata display standard errors based on Taylor series linearization, corresponding 95% confidence intervals for the percentage, and percentages rather than proportions, respectively):

```
. svy: tabulate dizzyrfv if subc == 1, se ci percent
```

Then we performed the same analysis, only using Stata's `subpop()` option for performing correct unconditional analyses (available for all `svy:` commands in Stata 10):

```
. svy, subpop(subc): tabulate dizzyrfv, se ci percent
```

Table 1 compares results from these two distinct analysis approaches.

Table 1. Results from the motivating example

	Conditional approach	Unconditional approach
Sample size	397	68,372
Subpopulation	397	397
Design strata	8	8
Design clusters	114	294
Design DF	106	286
Estimated percentage	4.8201	4.8201
Linearized standard error	1.5761	1.5904
95% confidence interval	(1.6954, 7.9448)	(1.6897, 7.9504)

In table 1, we note how the conditional approach resulted in substantially fewer design clusters being recognized by Stata (and therefore fewer design degrees of freedom being used for confidence-interval construction). When following this approach, Stata would not know that there are another 180 clusters that were a part of the 2004 NHAMCS design. Further, the conditional approach resulted in a smaller linearized standard error for the percentage, which in turn leads to a narrower confidence interval. While the resulting estimates and standard errors in this case are not substantially different when following these two approaches (we would not expect the actual weighted estimate to differ at all), the implications of following the (incorrect) conditional approach are apparent: underestimated standard errors and confidence intervals for statistics of interest that are too narrow (i.e., overstated precision of survey estimates). This problem tends to be exacerbated in complex samples with fewer PSUs selected from each sampling stratum.

Readers should keep in mind that the standard errors of these estimates express the degree of variability that we would expect from one hypothetical sample to another

around the true subpopulation percentage. In the conditional analysis, the calculations assumed that each hypothetical sample would have the same number of subpopulation visits (397) and all variation from sample to sample would come from variation in the value of the indicator for dizziness/vertigo among those subpopulation visits. In reality, this would never be the case because the subpopulation sample size is a random variable. The unconditional analysis accounts for this random variability.

4 Calculation of degrees of freedom for test statistics from subpopulations

The standard degrees-of-freedom calculation method in the analysis of complex-sample survey data is to assume that the design degrees of freedom are equal to the number of first-stage PSUs minus the number of design strata. The motivating example in table 1 shows that Stata uses this calculation in both the conditional and unconditional approaches. The appropriateness of this "fixed" degrees-of-freedom calculation method with respect to subpopulation analysis has recently been assessed for a subpopulation mean. Burns et al. (2003) considered a simulation with a true null hypothesis and showed that, when considering rare subpopulations not well-represented across a complex sample design, the use of this "fixed" method to calculate degrees of freedom may result in overestimation of degrees of freedom (and narrower confidence intervals). The consequence of this overestimation was an increased null hypothesis rejection rate, compared with a "variable" degrees of freedom calculation method involving only those sampled PSUs with observations in the subpopulation minus those strata with sampled observations in the subpopulation (Korn and Graubard 1999). Applying this "variable" method to the unconditional approach in section 3 would result in 106 design degrees of freedom for confidence-interval construction, rather than the full 286 degrees of freedom. We acknowledge that additional work is needed in this area.

Qian (1998) proposed additional improvements to the Satterthwaite method for the estimation of degrees of freedom for complex-sample survey data; these improvements built on work by Johnson and Rust (1992), which showed that Satterthwaite estimates tend to underestimate the effective degrees of freedom based on a complex design (resulting in conservative tests and wider confidence intervals). Rust and Rao (1996) offered practical guidance on calculating degrees of freedom for test statistics when working with replicated variance-estimation methods (see section 5). In short, they suggest that a working rule of thumb for approximating the effective degrees of freedom is to subtract the number of design strata with one or more subpopulation elements from the number of PSUs with one or more subpopulation elements (this is the "variable" method discussed above).

The unconditional variance-estimation approach discussed in this article emphasizes the use of all available strata and PSUs based on the complex design, and the full complex design is therefore reflected in the variance estimation equation for subpopulation statistics. However, *sampling zeros* occasionally arise in subpopulation analyses, where in a given sample a subpopulation is not represented in a given stratum (although it could

be). In this case, these strata should be omitted from variance estimation or degrees-of-freedom calculations when following the unconditional approach (Rust and Rao 1996). In Stata 10, the current versions of the Stata procedures for analysis of complex-sample survey data correctly drop these strata from degrees-of-freedom calculations. This was not an issue in the motivating example in section 3, but this is more likely to be an issue in complex designs with a greater number of sampling strata.

Accurate, robust determination by survey data analysts of degrees of freedom is needed both in the calculation of confidence intervals and the calculation of p-values for Wald-type statistics, making their proper determination essential for any tests of significance. Overstating the degrees of freedom for a given subpopulation analysis will tend to make confidence intervals too small, and this problem often arises when sampling zeros and *structural zeros* (strata with no subpopulation elements by design) are not correctly taken into account. Very little analytic theory exists for degrees-of-freedom calculations based on complex-sample designs, especially for subpopulations that are particularly small and "sparse" (i.e., not well-represented across the complex design).

5 Discussion

In the motivating example presented in section 3, we focused on the estimation of a percentage (and its corresponding standard error) for a specific subpopulation in a real survey dataset, collected in a medical care setting from a sample with a complex design. The issues discussed in this article also apply to estimates of means, totals, regression coefficients, odds ratios, and additional survey statistics of interest. In addition, we considered only Taylor series linearization for variance estimation in the examples that we presented. When analysts use replicated methods for variance estimation, such as jackknife repeated replication or balanced repeated replication, a conditional approach to variance estimation will result in correct standard errors, but the correct calculation of replicate weights representing the original complex design of the sample is essential. Most public-use datasets containing replicate weights have these weights calculated correctly by the agencies producing the datasets, so analysts do not need to worry about this weighting issue in practice. Determining the correct degrees of freedom for developing confidence intervals and conducting hypothesis tests when calculating replicated variance estimates for subpopulations is still an open problem. For more information on inference for subpopulations under jackknife repeated replication or balanced repeated replication, we refer readers to Rust and Rao (1996).

As indicated in this article, a conditional approach to the analysis of subpopulation data from a complex-sample survey restricts the estimation of sample statistics and variances only to those cases that belong to the defined subpopulation. Estimates of sampling variance and determination of degrees of freedom for confidence interval formation or hypothesis tests are "conditioned" on the strata and clusters in which, by design, sample observations actually appear. Theoretically, there are two situations in which the conditional approach to estimation and inference can be applied to subpopulation analysis of complex-sample survey data: 1) the subpopulation is a *design domain*

(Kish 1987), and observations on subpopulation members can be obtained only in a subset of design strata; and 2) the subpopulation size is known for each stratum, and appropriate stratified estimators are employed. Applying the unconditional analysis approach in most cases will yield correct analysis results for design domains, provided that sampling strata where it is not possible for a subpopulation to appear are not included in the calculation of degrees of freedom (see section 4).

The presence of missing data on analysis variables can effectively result in a conditional analysis, because cases with missing data on analysis variables will be deleted when running any types of analyses in nearly all statistical software packages. Analysts need to investigate missing-data problems carefully when performing any forms of complex-sample survey data analysis. Approaches to the analysis of missing data (such as multiple-imputation analysis) should be considered by analysts first to avoid the potential pitfalls associated with missing data in subpopulations of interest. We refer readers to Horton and Kleinman (2007) for additional alternatives to handling missing-data problems.

Survey data analysts can perform the unconditional subpopulation analysis presented in this article by using procedures in the SAS, SUDAAN, and SPSS software packages as well, and interested readers can email the authors for relevant code and options in these packages. We urge analysts to pay attention to the calculation of degrees of freedom according to a complex design (section 4) when performing subpopulation analyses in these other packages, taking the points that we have presented in this article into consideration (especially when working with sampling zeros).

Analysts should also keep in mind that performing conditional subpopulation analyses of complex-sample survey data can lead to problems where certain sampling strata have only a single PSU present, especially in paired-selection designs (where each sampling stratum has two PSUs selected). Procedures for survey data analysis in Stata will report an explicit note about this problem when it is encountered. Referring to (1) for the variance of sample totals, a more restrictive conditional approach to the analysis may result in a_h being equal to one or zero (meaning that stratum h has only a single PSU present in the restricted dataset), which would make the variance of the total undefined. The problem of having a single PSU per stratum (which prevents variance estimation) can arise in other situations in practice as well, and original designs with a single PSU selected from specific strata often need to be approximated with sampling-error calculations models that enable variance estimation. Interested readers can refer to Lee and Forthofer (2005, 42–43) or Korn and Graubard (1999, 140–141) for more information on this issue.

Finally, additional research is necessary to determine the most appropriate methods for calculating degrees of freedom for test statistics when performing subpopulation analyses. Simulations might be helpful in this area to assess the true distributions of test statistics, especially in the case of small subpopulations. The problem of analyzing extremely small subpopulations that theoretically could be represented across a full complex design but inevitably are not in any given sample also warrants future research; for these cases, there is a lack of analytic theory that does not rely on asymptotic results.

Model-based approaches to small-area estimation are currently recommended in practice (Rao 2005), and additional work is necessary to examine design-based approaches for small subpopulations.

6 References

Burns, A. M., R. J. Morris, J. Liu, and M. Z. Byron. 2003. Estimating degrees of freedom for data from complex surveys. In *Proceedings of the Survey Research Methods Section, American Statistical Association (2003)*. http://www.amstat.org/sections/srms/proceedings/y2003f.html.

Burt, C. W., and S. M. Schappert. 2004. Ambulatory care visits to physician offices, hospital outpatient departments, and emergency departments: United States, 1999–2000. *Vital and Health Statistics* 157: 1–70. http://www.cdc.gov/NCHS/data/series/sr_13/sr13_157.pdf.

Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. New York: Wiley.

Eltinge, J. L., and W. M. Sribney. 1996a. svy1: Some basic concepts for design-based analysis of complex survey data. *Stata Technical Bulletin* 31: 3–6. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 208–213. College Station, TX: Stata Press.

———. 1996b. svy2: Estimation of means, totals, ratios, and proportions for survey data. *Stata Technical Bulletin* 31: 6–23. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 213–235. College Station, TX: Stata Press.

———. 1996c. svy3: Describing survey data: Sampling design and missing data. *Stata Technical Bulletin* 31: 23–26. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 235–239. College Station, TX: Stata Press.

———. 1996d. svy4: Linear, logistic, and probit regressions for survey data. *Stata Technical Bulletin* 31: 26–31. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 239–245. College Station, TX: Stata Press.

———. 1996e. svy5: Estimates of linear combinations and hypothesis tests for survey data. *Stata Technical Bulletin* 31: 31–42. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 246–259. College Station, TX: Stata Press.

Fuller, W. A., W. Kennedy, D. Schnell, G. Sullivan, and H. J. Park. 1989. *CARP*. Ames, IA: Statistical Laboratory, Iowa State University.

Hing, E., S. Gousen, I. Shimizu, and C. Burt. 2003. Guide to using masked design variables to estimate standard errors in public use files of the National Ambulatory Medical Care Survey and the National Hospital Ambulatory Medical Care Survey. *Inquiry* 40: 401–415. http://www.cdc.gov/nchs/data/ahcd/ultimatecluster.pdf.

Horton, N. J., and K. P. Kleinman. 2007. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistician* 61: 79–90.

Johnson, E., and K. Rust. 1992. Effective degrees of freedom for variance estimates from a complex sample survey. In *Proceedings of the Survey Research Methods Section, American Statistical Association*.

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

———. 1987. *Statistical Design for Research*. New York: Wiley.

Korn, E. L., and B. I. Graubard. 1999. *Analysis of Health Surveys*. New York: Wiley.

Kreuter, F., and R. Valliant. 2007. A survey on survey statistics: What is done and can be done in Stata. *Stata Journal* 7: 1–21.

Lee, E. S., and R. N. Forthofer. 2005. *Analyzing Complex Survey Data*. 2nd ed. Thousand Oaks, CA: Sage.

Lohr, S. L. 1999. *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury.

McCaig, L. F., and T. McLemore. 1994. Plan and operation of the National Hospital Ambulatory Medical Survey. *Vital and Health Statistics* 34: 1–78. http://www.cdc.gov/nchs/data/series/sr_01/sr01_034acc.pdf.

Qian, J. 1998. Estimation of the effective degrees of freedom in t-type tests for complex data. In *Proceedings of the Survey Research Methods Section, American Statistical Association (1998)*. http://www.amstat.org/sections/srms/proceedings/papers/1998_119.pdf.

Rao, J. N. K. 2003. *Small Area Estimation*. Hoboken, NJ: Wiley.

———. 2005. Interplay between sample survey theory and practice: An appraisal. *Survey Methodology* 31: 117–138.

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

StataCorp. 2007. *Stata 10 Survey Data Reference Manual*. College Station, TX: Stata Press.

Wolter, K. M. 1985. *Introduction to Variance Estimation*. New York: Springer.

About the authors

Brady T. West is a lead statistician at the Center for Statistical Consultation and Research on the University of Michigan, Ann Arbor campus. He received an MA in applied statistics from the University of Michigan, and he will be joining the Survey Methodology Program as a PhD student in 2008. His primary research interests revolve around regression models for clustered and longitudinal data, and he has authored a book comparing different statistical software packages in terms of their mixed modeling procedures (*Linear Mixed Models: A Practical Guide Using Statistical Software*, Chapman Hall/CRC Press, 2007). He specializes in applications of statistical software and analysis of survey data. Through the Center for Statistical Consultation and Research, he teaches several yearly short courses on statistical methodology and software.

Patricia Berglund is a senior research associate in the Survey Methodology Program at the Institute for Social Research (ISR) on the University of Michigan, Ann Arbor campus. She holds an MBA from Northwestern University in Chicago, Illinois, and has extensive experience in the use of computing systems for data management and analysis. She is currently working in the mental health field, using data from the National Comorbidity Surveys, World Mental Health Surveys, and various other national and international surveys. In addition, she is involved in developing, implementing, and teaching analysis courses and computer training programs at the Survey Research Center/ISR.

Steven G. Heeringa is a research scientist in the Survey Methodology Program, the director of the Statistical and Research Design Group in the Survey Research Center, and the director of the Summer Institute in Survey Research Techniques at the Institute for Social Research on the University of Michigan, Ann Arbor campus. He has over 25 years of statistical sampling experience directing the development of the Survey Research Center National Sample design, as well as sample designs for the Survey Research Center's major longitudinal and cross-sectional survey programs. During this period, he has been actively involved in research and publication on sample design methods and procedures such as weighting, variance estimation, and the imputation of missing data that are required in the analysis of sample survey data. He has been a teacher of survey sampling methods to U.S. and international students and has served as a sample design consultant to a wide variety of international research programs based in countries such as Russia, Ukraine, Uzbekistan, Kazakhstan, India, Nepal, China, Egypt, Iran, and Chile.