Important Considerations When Analyzing Health Survey Data Collected Using a Complex Sample Design

Researchers often use survey data to answer important public health policy questions. Examples of common data sources used in public health research include the National Health and Nutrition Examination Survey, the National Health Interview Survey, the Medical Expenditure Panel Survey, and the National Ambulatory Medical Care Survey. All these surveys employ a complex sample design to recruit participants into the survey. When performing secondary analyses of complex sample survey data, it is necessary to remind ourselves of the key features of these designs that must be taken into account to produce valid statistical estimates.

There are (at least) 3 design features that are common to complex sample designs: (1) stratifying the population into nonoverlapping administrative areas (e.g., regions, states) within which independent samples are drawn; (2) sampling clusters (or groups) of individual units instead of directly sampling individual units, as in a simple random sample; and (3) assigning unequal probabilities of selection to individual units in the population, often with the goal of oversampling certain subgroups (e.g., minority racial/ethnic groups). Failure to account for these design features can lead to biased estimates (or statistics), incorrect variance estimates and confidence intervals that are too narrow, and, most importantly, misleading conclusions. We discuss each of these features in turn and explain how to identify and account for them when analyzing complex sample survey data.

The first design feature, stratification, involves dividing the entire population into nonoverlapping areas. The areas are divided such that the characteristics of the population are homogenous within areas and heterogeneous between areas. Examples of strata include geographical regions or states. Independent samples are drawn from each stratum to ensure that each area is represented in the sample. Stratification yields greater precision and smaller standard errors for statistical estimates computed from the survey data because of the removal of between-stratum variance from the sampling variance of survey estimates. Survey analysis software can incorporate variables indicating stratification codes (e.g., STRATUM) into the analysis to produce more precise statistical estimates.

The second design feature, cluster sampling, involves sampling groups of individual units rather than the units themselves (as in the case of simple random sampling). Examples of clusters include hospitals, schools or classrooms, and neighborhoods. In a complex sample design, clusters are usually sampled within strata. Although cluster sampling is a cost-effective method for administering interviews (relative to simple random sampling), it has the disadvantage that units within a cluster often resemble each other in the characteristics of interest. These intracluster correlations generally result in decreased precision and larger standard errors for statistical estimates computed from the survey data. If this additional variation is not accounted for in the analysis, standard errors will be underestimated and point estimates may be incorrectly interpreted as being statistically significant ($P<.05$) when in actuality they are not. When provided, cluster identifiers usually take the name PSU or CLUSTER.

The third design feature common to complex sample surveys is that some individual units are selected into the sample with a higher probability than are others. This is typically done to increase the sample size for minority groups in the population. For example, researchers may want to compute statistical estimates for a particular subgroup, say, Hispanics. Because Hispanics make up approximately 17% of the US population, an equal probability sample of the population would result in about 17% of the sample consisting of Hispanics, on average. The size of this sample may be too small to compute reliable design-based estimates for this subgroup. Alternatively, most surveys would assign Hispanics a higher probability of selection that would result in a larger number of Hispanics selected into the sample. However, it is important to note that the composition of the sample would no longer resemble the composition of the population, and the resulting statistical estimates would be biased in favor of a larger Hispanic population if the unequal selection probabilities are not accounted for.

Complex sample surveys produce sampling weights that researchers can use to account for the unequal probabilities of selection in their estimates. This is especially important for descriptive estimates; weights may or may not affect inferences about the coefficients in regression models.[1] The sampling weight variable is usually listed under the name WEIGHT. The sampling weights may also be adjusted for unit nonresponse and population noncoverage. The purpose of these additional adjustments is to reduce the bias associated with these sources of nonobservation error. Some survey data sets may include replicate weights, enabling appropriate variance estimation without the need to include stratum and cluster identifiers (which is important when disclosure risk is a concern).

Accounting for these design features when computing population estimates from complex sample survey data is straightforward using specialized statistical software. Commonly used statistical software packages, including R (R Foundation for Statistical Computing, Vienna, Austria), Stata (StataCorp LP, College Station, TX), SAS (SAS Institute, Cary, NC), and SPSS (SPSS, Inc, Chicago, IL), have procedures that incorporate the stratification, clustering, and sampling weight variables into the statistical estimation process. For survey analysis in R, the "survey" package[2] is recommended because it offers many facilities for conducting analyses of complex sample survey data. For Stata, the survey (SVY) commands can be used to account for any and all complex design features for a variety of statistical estimators.[3] In SAS, a variety of built-in procedures are available for univariate and multivariate analyses. The procedures are usually prefaced by the word SURVEY (e.g., SURVEYMEANS, SURVEYREG). The Complex Samples add-on module in SPSS also permits the computation of valid statistical estimates from complex sample survey data.

Ultimately, it is the responsibility of the researcher to identify the relevant design features in complex sample surveys and to employ appropriate statistical estimation methods that account for these features when producing estimates from such surveys. Survey organizations spend considerable sums of money to design, collect, and disseminate data from complex sample surveys, and these costly efforts would go to waste if users of these data fail to recognize the sample design features that give rise to the samples studied by health researchers. ■

Joseph W. Sakshaug, PhD
Brady T. West, PhD

About the Authors

Joseph W. Sakshaug is with the Department of Statistical Methods, German Institute for Employment Research, Nuremberg, and the Ludwig Maximilian University of Munich, Munich, Germany. Brady T. West is with the Survey Research Center and the Center for Statistical Consultation and Research, University of Michigan, Ann Arbor.

Correspondence should be sent to Joseph W. Sakshaug, Department of Statistical Methods, Institute for Employment Research, 104 Regensburger Strasse, Nuremberg 90478, Germany (e-mail: joesaks@umich.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This article was accepted June 16, 2013.
doi:10.2105/AJPH.2013.301515

Contributors

Both authors contributed to the writing of the article.

Acknowledgments

J. W. Sakshaug was supported by the Alexander von Humboldt Foundation.

References

1. Korn EL, Graubard BI. *Analysis of Health Surveys.* New York, NY: Wiley; 1999.

2. Lumley T. *Complex Surveys: A Guide to Analysis Using R.* Hoboken, NJ: John Wiley & Sons; 2010.

3. Heeringa SG, West BT, Berglund PA. *Applied Survey Data Analysis.* Boca Raton, FL: Chapman & Hall; 2011.