

# Analysis of Complex Survey Data

Brady West, Ph.D.

*Upcoming Seminar:*  
May 6-8, 2021, Remote Seminar

## Recommended Readings

- Cochran, W.G. *Sampling Techniques*, 3<sup>rd</sup> Edition (1977). New York: John Wiley & Sons. (*Mainly Theoretical*)
- Chambers, R.L., and Skinner, C.J. (editors). (2003). *Analysis of Survey Data*. New York: John Wiley & Sons. (*Very Theoretical*)
- Heeringa, S.G., West, B.T., and Berglund, P.A. (2017). *Applied Survey Data Analysis, Second Edition*. Chapman & Hall / CRC Press, Boca Raton, FL. (*Practical / Stata*)**
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons. (*Both*)
- Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*. John Wiley and Sons, NY. (*Both*)
- Lee, E.S. and Forthofer, R.N. (2006). *Analyzing Complex Sample Survey Data* (2<sup>nd</sup> Edition). Sage, Thousand Oaks, CA. (*Practical*)
- Lohr, S. L. (1999). *Sampling: Design and Analysis*, Duxbury Press, Pacific Grove, CA. (*Both*)
- Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons, Hoboken, NJ. (*Practical*)
- Skinner, C.J., Holt, D., Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons. (*Mainly Theoretical*)
- Wolter, K.M. *Introduction to Variance Estimation*, 2<sup>nd</sup> Edition (2007). New York: Springer-Verlag. (*Mainly Theoretical*)

# Seminar Objectives

- Develop an understanding of design-based estimation and inference for complex sample survey data.
- Acquire the ability to apply the theory to a range of problems encountered in survey practice.
- Learn how to use software tools in R and Stata that are available for analysis of complex sample survey data.
- Gain an understanding of how to interpret the results of an analysis of complex sample survey data.



## Applied Survey Data Analysis

- [Project Overview](#)
- [Information about Authors](#)
- [Links to Data Sets](#)
- [Links to Additional Sites](#)
- [Survey Data Analysis Publications](#)
- [Professional Reviews](#)
- [Frequently Asked Questions](#)
- [Supplemental Code](#)

### Site Overview

This site contains information about the text *Applied Survey Data Analysis*, (first and second editions) including author biographies, links to public release data sets and related sites, code and output for analysis examples replicated in current software packages, and information about new publications of interest to survey data analysts. Other features include a FAQ log and links to other software and statistical sites. We plan to intermittently update this site with news about ongoing statistical and software advances in the field of analysis of survey data.

### Special Notes from Authors

**ASDA-Second Edition is Available as of June 28, 2017!**

### Project Overview

*Applied Survey Data Analysis* is the product born of many years of teaching applied survey data analysis classes and practical experience analyzing survey data. We have taught various versions of this course in the ISR/SRC Summer Institute Program, as part of University of Michigan/CSCAR, and within the Survey Methodology Program at University of Michigan and University of Maryland. Our goal has been to integrate teaching materials and practical analysis knowledge into a textbook geared to a level accessible for graduate students and working analysts who may have varying levels of statistical and analytic expertise. We intend to update the materials on this website as statistical and software improvements emerge with the goal of assisting analyst and researchers performing survey data analysis.

### Information About Authors

**Patricia A. Berglund** is a Senior Research Associate in the Survey Methodology Program at the Institute for Social Research. She has extensive experience in the use of computing systems for data management and complex sample survey data analysis. She works on research projects in youth substance abuse, adult mental health, and survey methodology using data from Army STARRS, Monitoring the Future, the National Comorbidity Surveys, World Mental Health Surveys, Collaborative Psychiatric Epidemiology Surveys, and various other national and international surveys. In addition, she is involved in development, implementation, and teaching of analysis courses and computer training programs at the Survey Research Center-Institute for Social Research. She also lectures in the SAS@ Institute-Business Knowledge Series. <mailto:pberg@umich.edu>

**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. He has over 25 years of statistical sampling experience directing the development of the SRC National Sample design, as well as sample designs for SRC'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, the Ukraine, Uzbekistan, Kazakhstan, India, Nepal, China, Egypt, Iran, and Chile. <mailto:sheering@umich.edu>

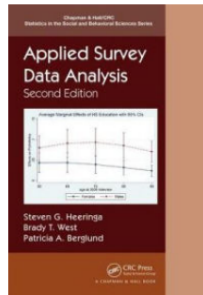
**Brady T. West** Brady T. West is a Research Associate Professor in the Survey Methodology Program, located within the Survey Research Center at the Institute for Social Research on the University of Michigan-Ann Arbor (U-M) campus. He also serves as a Statistical Consultant on the U-M Consulting for Statistics, Computing, and Analytics Research (CSCAR) team. He earned his PhD from the Michigan Program in Survey Methodology in 2011. Before that, he received an MA in Applied Statistics from the U-M Statistics Department in 2002, being recognized as an Outstanding First-year Applied Masters student, and a BS in Statistics with Highest Honors and Highest Distinction from the U-M Statistics Department in 2001. His current research interests include the implications of measurement error in auxiliary variables and survey paradata for survey estimation, survey nonresponse, interviewer variance, and multilevel regression models for clustered and longitudinal data. He is the lead author of a book comparing different statistical software packages in terms of their mixed-effects modeling procedures (Linear Mixed Models: A Practical Guide using Statistical Software, Second Edition, Chapman Hall/CRC Press, 2014), and he is a co-author of a second book entitled Applied Survey Data Analysis (with Steven Heeringa and Pat Berglund), which was published by Chapman Hall in April 2010 and has a second edition in press that will be available in mid-2017. Brady lives in Dexter, MI with his wife Laura, his son Carter, his daughter Everleigh, and his American Cocker Spaniel Bailey. <mailto:bwest@umich.edu>

### Professional Reviews of ASDA-Second Edition

Review/Summary from Chapman Hall Website

#### Features

- Bootstrap methods of variance estimation.
- Estimation and inference for specialized functions such as the Gini coefficient and log-linear models.



### Analysis Examples Replication-Second Edition

The analysis examples replication materials cover Chapters 5-13 of ASDA Second Edition but not every software package contains all chapters. Lack of a link for a given chapter indicates that this software package does not include the ability to perform this type of analysis technique.

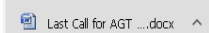
### SAS v9.4 Code and Results

[Overview of SAS Commands](#)

[Chapter 5 Analysis Examples](#)

[Chapter 6 Analysis Examples](#)

[Chapter 7 Analysis Examples](#)



# ASDA Web Site - Example Code and Results

The screenshot shows a web browser window with the URL [isr.umich.edu/src/smp/asda/](http://isr.umich.edu/src/smp/asda/). The page content is organized into several sections:

- SAS v9.4 Code and Results**
  - [Overview of SAS Commands](#)
  - [Chapter 5 Analysis Examples](#)
  - [Chapter 6 Analysis Examples](#)
  - [Chapter 7 Analysis Examples](#)
  - [Chapter 8 Analysis Examples](#)
  - [Chapter 9 Analysis Examples](#)
  - [Chapter 10 Analysis Examples](#)
  - [Chapter 11 Analysis Examples](#)
  - [Chapter 12 Analysis Examples](#)
- Stata v14 Code (Results Are Presented Throughout the Book)**
  - [Overview of Stata Commands](#)
  - [Chapter 5 Analysis Examples](#)
  - [Chapter 6 Analysis Examples](#)
  - [Chapter 7 Analysis Examples](#)
  - [Chapter 8 Analysis Examples](#)
  - [Chapter 9 Analysis Examples](#)
  - [Chapter 10 Analysis Examples](#)
  - [Chapter 11 Analysis Examples](#)
  - [Chapter 12 Analysis Examples](#)
  - [Chapter 13 Analysis Examples](#)
- SPSS V22 Code and Results**
  - [Overview of SPSS Commands](#)
  - [Chapter 5 Analysis Examples](#)
  - [Chapter 6 Analysis Examples](#)
  - [Chapter 7 Analysis Examples](#)
  - [Chapter 8 Analysis Examples](#)
- Professional Reviews of ASDA-Second Edition**
  - Review/Summary from Chapman Hall Website**
    - Features**
      - Bootstrap methods of variance estimation.
      - Estimation and inference for specialized functions such as the Gini coefficient and log-linear models.
      - Updated approaches to examining model diagnostics, testing goodness of fit, and estimation and display of marginal effects in linear and generalized linear models.
      - State-of-the-art methods for analysis of longitudinal survey data.
      - Fractional imputation methods for item missing data.
      - Enhanced treatment of methods and software for fitting multilevel models, structural equation models and other latent variable models to complex sample survey data.
      - Updated review of software packages for the analysis of complex sample survey data.
    - Summary**

Highly recommended by the Journal of Official Statistics, The American Statistician, and other journals, Applied Survey Data Analysis, Second Edition provides an up-to-date overview of state-of-the-art approaches to the analysis of complex sample survey data. Building on the wealth of material on practical approaches to descriptive analysis and regression modeling from the first edition, this second edition expands the topics covered and presents more step-by-step examples of modern approaches to the analysis of survey data using the newest statistical software.

Designed for readers working in a wide array of disciplines who use survey data in their work, this book continues to provide a useful framework for integrating more in-depth studies of the theory and methods of survey data analysis. An example-driven guide to the applied statistical analysis and interpretation of survey data, the second edition contains many new examples and practical exercises based on recent versions of real-world survey data sets. Although the authors continue to use Stata for most examples in the text, they also continue to offer SAS, SPSS, SUDAAN, R, WesVar, IVEware, and Mplus software code for replicating the examples on the book's updated Web site.
  - Links to Data Sets for First and Second Editions**
    - National Comorbidity Survey-Replication (Collaborative Psychiatric Epidemiology Surveys)**
      - <http://www.icpsr.umich.edu/cpes> (for online documentation tools and data download)
      - <http://www.hcp.med.harvard.edu/hcs> (for NCS-R specific information)
    - National Health and Nutrition Examination Survey (National Center for Health Statistics)**
      - <http://www.cdc.gov/nchs/>
    - Health and Retirement Survey (Institute for Social Research-University of Michigan)**
      - <http://hrsonline.isr.umich.edu>
    - European Social Survey (ESS)**
      - <http://www.europeansocialsurvey.org/>
    - United States Census Bureau**
      - <http://www.census.gov/>
  - Chapter Exercises Data Sets - Second Edition**

# Getting the Seminar Data Sets into R and Stata

- We will be working with a variety of survey data sets from various large-scale survey data collection programs, such as the European Social Survey (ESS) and the National Health and Nutrition Examination Survey (NHANES)
- **Example R Code for reading in the ESS-Russia Data:**

```
# install tidyverse packages  
install.packages("tidyverse")
```

```
# load tidyverse packages  
library("tidyverse")
```

```
# create russia data frame object based on ESS-Russia data  
russia <- readr::read_csv('http://umich.edu/~bwest/ess_russia.csv')
```

- **Example Stata Code for reading in the NHANES Data:**

```
use "http://www-personal.umich.edu/~bwest/nhanes1112_sub_10jun2016.dta",  
clear
```

# **INTRODUCTION: COMPLEX SAMPLES**

# Complex Sample Survey Data: Probability Samples

- **Probability sample design:**
  - Each population element has a known, non-zero selection probability
  - Properly weighted, sample estimates are unbiased or nearly unbiased for the corresponding population statistic
  - Variance of sample statistics can be estimated from the sample data (measurability)
- **Simple random sample (SRS):** A probability sample in which each element has an independent and equal chance of being selected for observation. Closest population sampling analog to independently and identically distributed (iid) data.



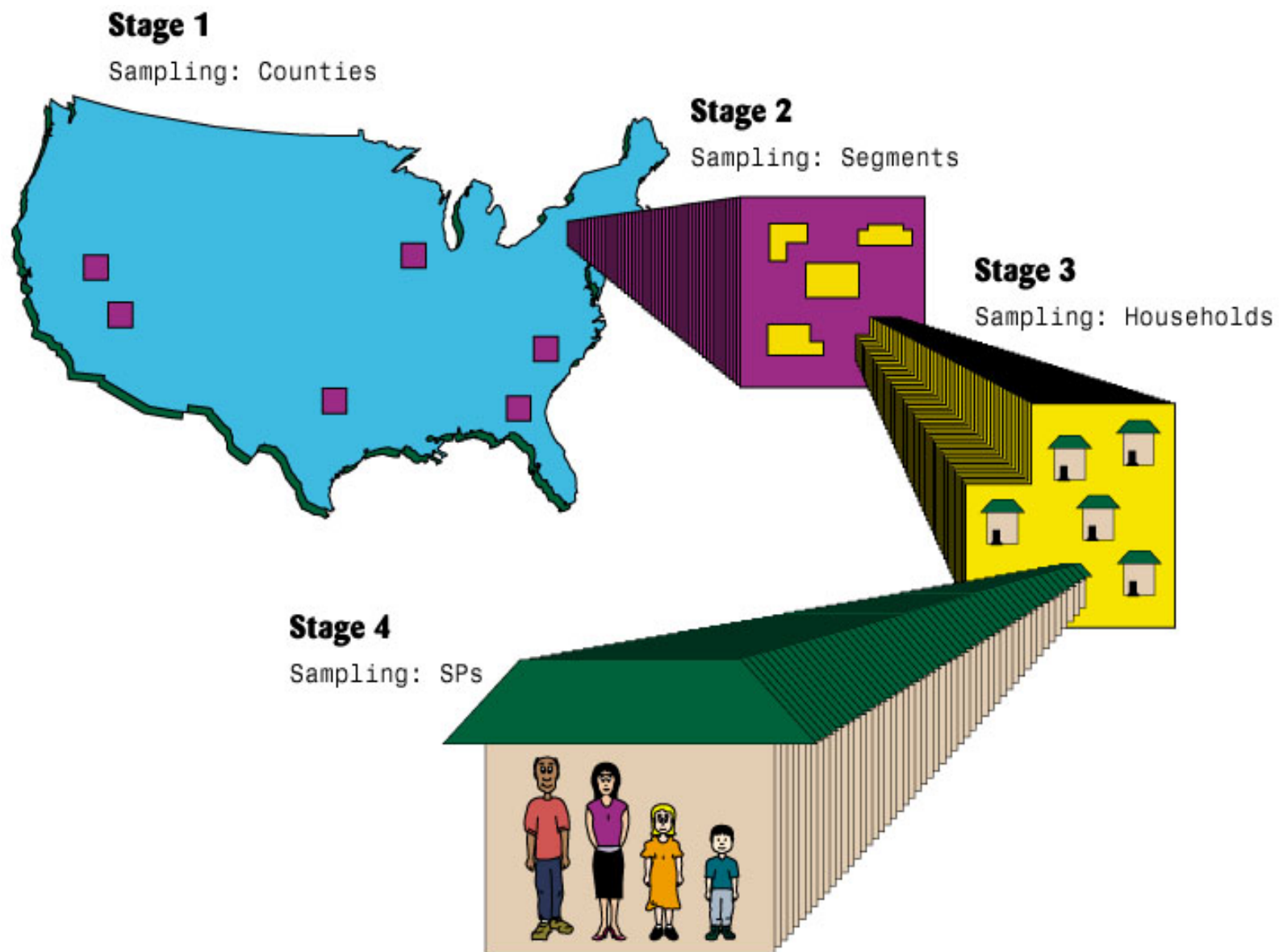
# Complex Sample Survey Data: “Complex” Designs

- “Complex sample”:
  - A probability sample developed using sampling procedures such as stratification, clustering and weighting, and designed to improve statistical efficiency, reduce costs or improve precision for subgroup analyses relative to SRS
  - Unbiased estimates with measurable sampling error are still possible
  - Independence of observations, (iid), equal probabilities of selection may no longer hold

# Where are Complex Sample Designs Used?

- Complex sample designs are the rule and not the exception in sample-based studies in the Social Sciences, Epidemiology, Public Health, Agriculture, Natural Resources and many other scientific fields.
- Common designs
  - Area probability samples of household populations
  - Multi-stage samples of schools, classes and students
  - Stratified samples of businesses, hospitals
- Other designs
  - Dual-frame and other samples in agriculture
  - RDD telephone samples
  - Natural resource samples (time and location samples)

# Multi-Stage Sample of U.S. Households: Illustration



# NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES)

- NHANES I (1971-75)
- NHANES II (1976-80)
- NHANES III (1988-94)
  
- NHANES: 1999-present. NHANES is now a continuous survey. New replicate samples every year.
  - ASDA (Second Edition) uses the NHANES 2011-2012 data set.

# NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES)

- Sampled persons are interviewed about health and health-related matters and undergo medical examinations in mobile examination centers (MECs).
  - <http://www.cdc.gov/nchs/nhanes>
- Separate estimates required for domains defined by the cross-classification of:
  - Age-sex groups, African Americans, Mexican Americans, and All Others.
  - Adolescents, the elderly, Mexican Americans, and the African-American population are oversampled.