STATISTICAL HORIZONS

Analysis of Complex Survey Data

Brady West, Ph.D.

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

Recommended Readings

Cochran, W.G. *Sampling Techniques*, 3rd 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* (2nd 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*)

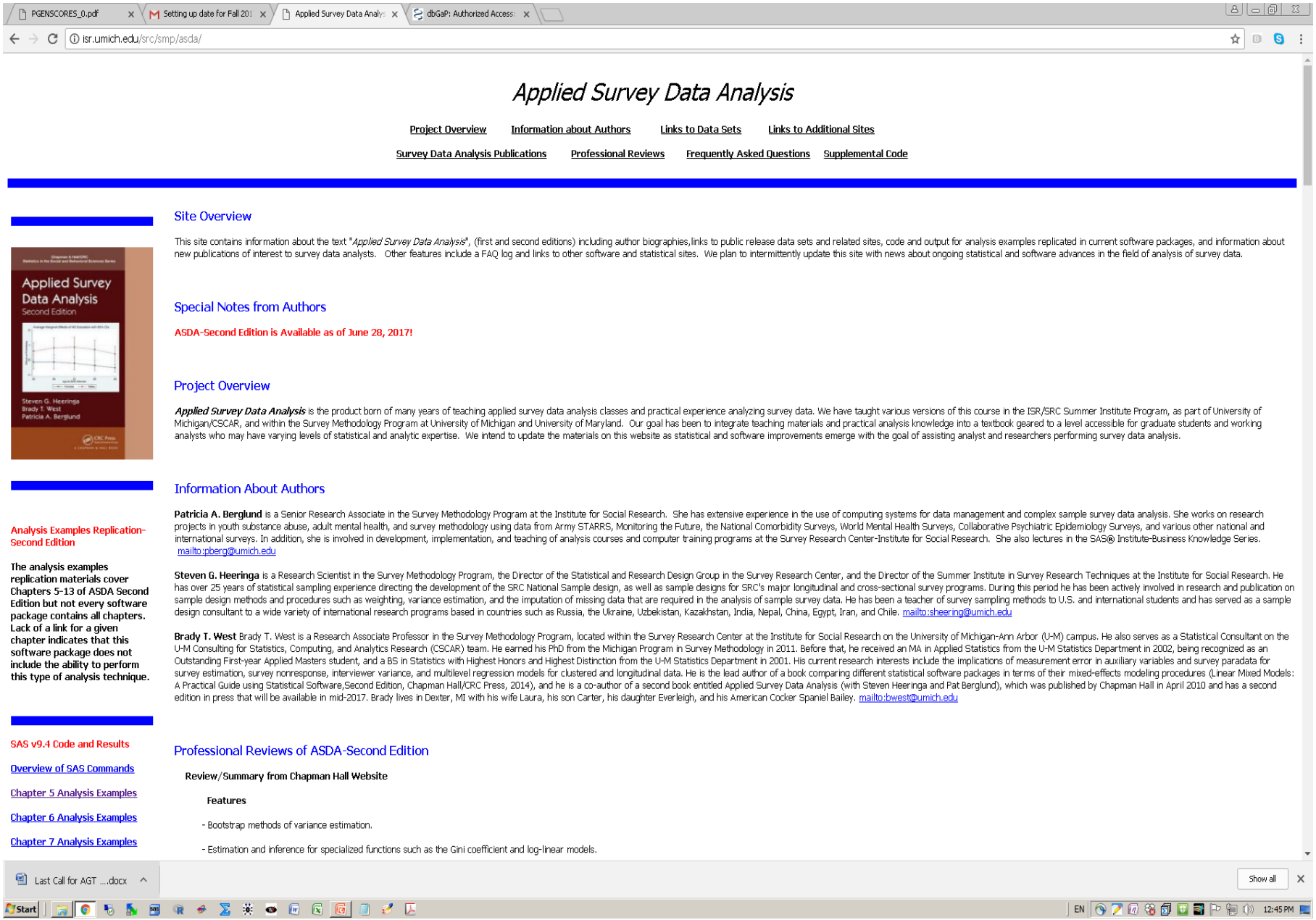
Skinner, C.J., Holt, D., Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons. (*Mainly Theoretical*)

Wolter, K.M. *Introduction to Variance Estimation*, 2nd Edition (2007). New York: Springer-Verlag. (*Mainly Theoretical*)

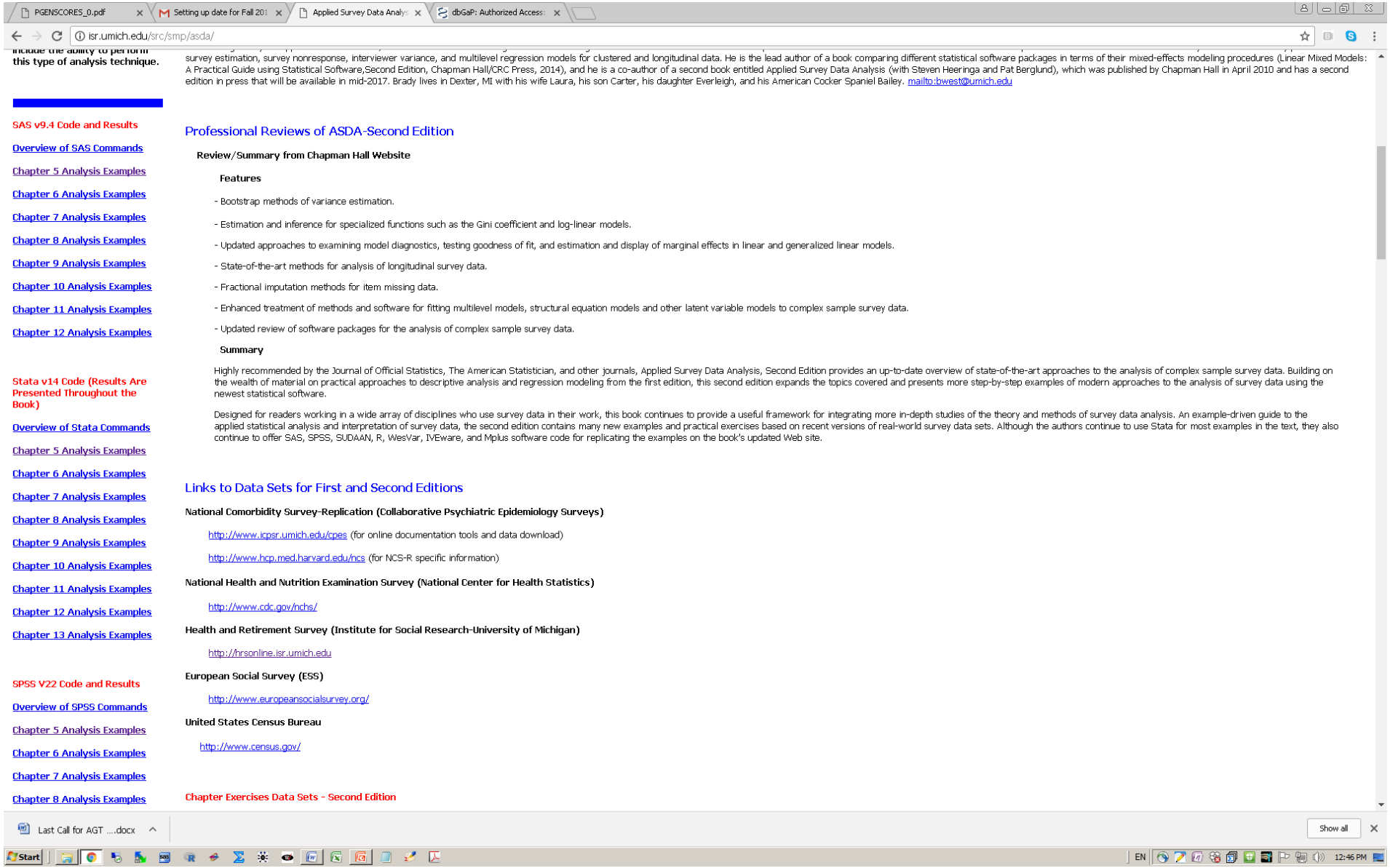
Seminar Objectives

- 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.

http://www.isr.umich.edu/src/smp/asda/



ASDA Web Site - Example Code and Results



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

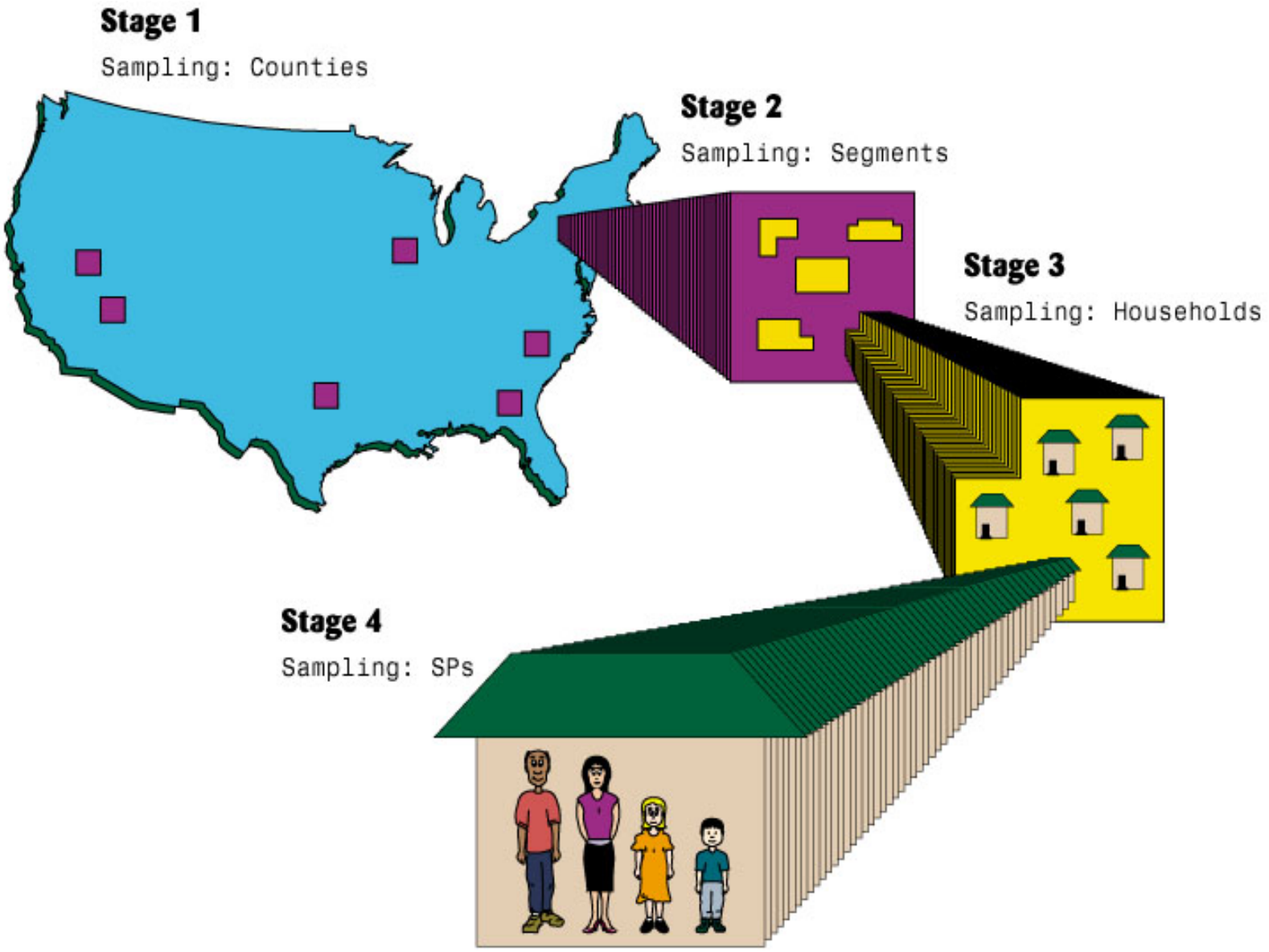
Where are Complex Sample Designs Used?

- 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.