Course Description and Syllabus

SURV 701 / SURVMETH 613
Analysis of Complex Sample Survey Data

Joint Program in Survey Methodology, University of Maryland
Program in Survey Methodology, University of Michigan

FALL, 2012

Abstract: Analysis of Complex Sample Survey Data covers the following topics: the development and handling of selection and other compensatory weights for survey data analysis; the effects of stratification and clustering on survey estimation and inference; alternative variance estimation procedures for estimated survey statistics; methods and computer software that take into account the effects of complex sample designs on survey estimation and inference; and methods for handling missing data, including weighting adjustment and imputation.

Instructor: Brady T. West, Ph.D.

Course: SURV 701 / SURVMETH 613
Dates: September 6 – December 13, 2012
Lectures: Thursday, 12:30pm - 3:00pm
Locations: University of Michigan, 300 Perry
University of Maryland, LeFrak Hall Room 1208
U.S. Census Bureau, Room T10

Brady T. West
4021 ISR
(734) 647-4615
bwest@umich.edu

Teaching Assistants and Course Advisors:

Pat Berglund (Advisor) Hanyu Sun (Grader)
pberg@isr.umich.edu hsun@survey.umd.edu

Course Description

Standard courses on statistical analysis assume that survey data are from a simple random sample of the target population. Little attention is given to characteristics often associated with survey data, including missing data, unequal probabilities of observation, and stratified multistage sample designs. Most standard statistical programs in software packages commonly used for data analysis (e.g., SAS®, SPSS®, and Stata®) do not allow the analyst to take most of
these properties of survey data into account. Failure to do so can have an important impact on
the estimation and inference for all types of analyses, ranging from simple descriptive statistics
to the estimation of parameters of multivariate models.

This course provides an introduction to procedures and software programs that have been
developed for the analysis of complex sample survey data. The course begins by considering the
sample designs of specific surveys: the National Comorbidity Survey Replication (NCS-R), the
2005-2006 National Health and Nutrition Examination Survey (NHANES), and the 2006 Health
and Retirement Study (HRS). Relevant design features of the NCS-R, NHANES and HRS
include weights that take into account differences in probability of selection into the sample and
differences in subgroup response rates, in addition to the stratification and cluster sampling
employed in the multistage sampling procedures used to select households and individuals.

The course will then move on to the introduction of variance estimation techniques that
have been developed to take into account the stratification and cluster sampling that are
properties of the multistage sampling designs used by most major survey programs. These will
initially be discussed in terms of the estimation of sampling variances for descriptive statistics:
sample means, proportions and quantiles of distributions. The course syllabus will then turn to
software and procedures for commonly used analyses, including testing for between-group
differences in means and proportions, linear regression analysis for continuous dependent
variables, contingency table analysis for categorical data and logistic regression for categorical
responses, generalized linear models for ordinal and count data, survival analysis and multilevel
modeling. We will also consider the consequences of nonresponse and missing data on survey
analysis and methods for dealing with missing data.

The SAS® and Stata® systems for data management and analysis will be used to develop
course examples and exercises. Illustrations will also be presented using the SUDAAN® and
WesVar® software systems that have been specifically designed for the analysis of survey data.
Data from the NCS-R, NHANES and HRS will be used to illustrate the various analysis
procedures covered during the course.

Prerequisites

The prerequisites for SURV 701 / SURVMETH 613 include one or more graduate
courses in statistics, a course in applied sampling methods, or permission of the instructor. The
course is presented at a moderately advanced statistical level. Although the course will review
the fundamentals of statistical analysis methods for survey data and provide detailed examples
on the use of statistical software, it will be assumed that the students are familiar with statistical
methods, including multiple regression and logistic regression. The initial lectures in the course
syllabus will review the various complex features of sample designs and how they influence
estimation and inference based on survey data. The course syllabus and level of instruction also
assume that students are familiar with basic sampling procedures, including simple random
sampling, stratification, cluster sampling and multi-stage sample designs. Students who do not
have graduate-level training in sampling techniques should expect to devote additional time
during the first weeks of the course to supplemental readings on this topic.
Course Format

Students in the class will meet in classrooms in Ann Arbor and College Park, and also classrooms at the United States Census Bureau and Bureau of Labor Statistics. These classrooms will be linked by an interactive video system that will allow the students in all four locations to see as well as hear the instructor and students in the other locations, and to view materials on an overhead display. The instructor will physically be in College Park for four of the class sessions, and in Ann Arbor for the remaining sessions.

Class time will be used for a combination of lectures and discussion of examples and homework. Lecture notes and examples will be presented on Power Point slides and copies of these materials will also be available to each student on the course web site (C-Tools). Questions are welcomed during lectures, and discussion of the topics is encouraged.

Textbook and Class Reading

The textbook for this course will be *Applied Survey Data Analysis* (ASDA, 2010; publisher: Chapman Hall / CRC Press), authored by the course instructors. Students can purchase the course text from online retailers (e.g., Amazon.com, or crcpress.com). Assigned readings will generally consist of selected sections from the chapters in the course text. The instructors also recommend that students who have a strong interest in the theory of analysis of complex sample survey data consider purchasing a copy of *Analysis of Survey Data*, edited by Chambers and Skinner (2003).

In addition to assigned readings from the course text (ASDA), the instructors have prepared a supplemental readings list that includes several review articles. These supplemental readings are provided in electronic format via the University of Michigan C-Tools system. Some of the supplemental readings on C-Tools will be assigned, and others will be recommended. **Students are required to have finished all assigned readings prior to the lecture for which they have been assigned.** Starting September 13, each lecture will begin at 12:40pm with a brief closed-book quiz that asks general questions about the readings.

Grading

The course grade will be based on four criteria:

- Weekly quizzes on assigned readings (10%);
- Completion of seven (7) homework assignments (30%);
- Take-home Mid-term exam (20%); and
- Final course project (40% of total grade).

Frequency of participation in weekly discussions will also be accounted for when determining final grades. The take-home mid-term exam will be distributed on **October 25**, and will involve a more extensive analysis of a selected survey data set.
Homework Assignments and Course Projects

The course homework assignments will be computational exercises that are provided by the instructor as a handout or based on exercises at the end of chapters of ASDA. These exercises generally involve analyzing a specified survey data set and then interpreting the results of those analyses. These analyses can be done on your own PC or workstation (if you have the necessary software) or on PCs provided for students at the respective sites. The minimum necessary software includes SAS® and Stata®. The data needed for these assignments will be provided to you via the course web site (C-Tools). Basic familiarity with the use of computers is assumed. Ability to work with SAS® or Stata® is also assumed, but students can easily learn the fundamentals as a part of the course exercises. Examples of the setups and interpretation of the output will be provided in class prior to assignments that require use of a particular procedure.

Students are encouraged to work in groups on the homework assignments. However, the work that is submitted must be done by each student; group submissions are not acceptable. Academic integrity as outlined in the respective graduate student policies at the two Universities will be required, and apparent violations of these policies will be dealt with on an individual basis. Assignments are due at noon on the specified dates, without exception, and the instructor will answer questions about the assignments in the C-Tools Chat Room from 7:00pm-8:00pm on the Tuesday before each due date. Assignments are required to be submitted via email to Dr. West (bwest@umich.edu) and Hanyu Sun (hsun@survey.umd.edu) in PDF format, with the filename having the following example structure: WEST_HW1_SM613.pdf / WEST_HW1_S701.pdf. A failure to follow this formatting (e.g., submitting .docx files or .xlsx files, or using incorrect file names) will result in the deduction of 1 point from the final assignment score. Assignments will each be graded on a 20-point basis, and grades of 0 will be given for any missed assignments. Assignments will be graded electronically and returned a week after their due dates, and the solutions will be discussed in class on that date.

Final Class Project

The primary aims of this course are to provide class participants with instruction in the theory and experience in the application of the software and methods for the analysis of complex sample survey data. The ultimate goal of this course is to prepare students to apply appropriate methods and software in the analysis of survey data and to effectively communicate the results of their analysis in the form of papers, technical reports or others forms of scientific communication. To this end, the course will require each student to develop a final project paper based on an independent analysis of a survey data set. The survey data set may be identified by the student or chosen from a list of course data sets. Work on the final project paper will begin in weeks 4 through 7 of the course, with a topic search and investigation of potential data sets. Selection of a project survey data set and topic will be finalized by Thursday, October 25. A preliminary draft of the final paper with the initial sections (background, literature review, data and methods) will be due Thursday, November 29. The final paper will be due to the instructors in electronic format by 5:00pm on Friday, December 14 (no exceptions). The instructors will be available throughout the course to assist students in each successive phase of the development of the final project paper.
### SURV 701 / SURVMETH 613: Analysis of Complex Sample Survey Data  
Fall 2012 Topics, Reading Assignments, and Homework

* With the exception of September 6, assigned readings need to be completed prior to the start of the indicated class, and the quiz will cover the assigned material.

<table>
<thead>
<tr>
<th>Class Date</th>
<th>Topic/Homework</th>
<th>Assigned Readings*</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 6</td>
<td>Survey estimation and inference for complex sample designs (Part 1). Multi-stage designs, stratification, cluster sampling, finite population corrections, design effects, effective sample size. Introduction to course data sets.</td>
<td>Syllabus</td>
</tr>
<tr>
<td></td>
<td><strong>Homework 1 Distributed.</strong></td>
<td>Chapters 1 and 2, ASDA</td>
</tr>
<tr>
<td>September 13</td>
<td>Survey estimation and inference for complex sample designs (Part 2). Weighting, models and assumptions for inference from complex sample data, sampling distributions, and confidence intervals.</td>
<td>Chapter 3 (3.1-3.5), ASDA</td>
</tr>
</tbody>
</table>
|                 | **CTOOLS:**  
|                 | 2. Juster and Suzman (1995)                                                                                                                                                                                  | **Homework 1 Due.** |
|                 | **Homework 1 Due.**  
|                 | **Homework 2 Distributed.**                                                                                                                                                                                  | **CTOOLS:**  
|                 | 1. Rust (1985)  
| (UMD)           | **Homework 2 Due.**  
|                 | **Homework 3 Distributed.**                                                                                                                                                                                  | **CTOOLS:**  
<p>|                 | Kreuter and Valliant (2007)                                                                                                                                                                                  |</p>
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Chapter/References</th>
</tr>
</thead>
<tbody>
<tr>
<td>October 11</td>
<td>Analysis Methods for Categorical Data. Homework 4 Due. Homework 5 Distributed. Final Project Introduced.</td>
<td>Chapter 6, ASDA</td>
</tr>
<tr>
<td>October 18</td>
<td>Linear Regression Review. Homework 5 Due.</td>
<td>Chapter 7 (7.1-7.3), ASDA</td>
</tr>
<tr>
<td>October 25</td>
<td>Linear Regression for Complex Samples. Final Project Prospectus Due. Take-Home Midterm Distributed.</td>
<td>Chapter 7 (7.4-7.5), ASDA</td>
</tr>
<tr>
<td>November 1</td>
<td>Logistic Regression for Complex Samples. Homework 6 Distributed. Take-Home Midterm Due.</td>
<td>Chapter 8, ASDA</td>
</tr>
<tr>
<td>November 8</td>
<td>Multinomial, Ordinal Logistic Regression. Homework 6 Due. Homework 7 Distributed.</td>
<td>Chapter 9 (9.1-9.3), ASDA</td>
</tr>
<tr>
<td>November 15</td>
<td>Poisson, Negative Binomial Regression. Homework 7 Due.</td>
<td>Chapter 9 (9.4), ASDA</td>
</tr>
<tr>
<td>December 6</td>
<td>Survival Analysis with Survey Data.</td>
<td>Chapter 10, ASDA</td>
</tr>
<tr>
<td>December 13</td>
<td>Multilevel Modeling of Survey Data.</td>
<td>Chapter 12, ASDA</td>
</tr>
<tr>
<td>December 14</td>
<td>Final Project Due.</td>
<td>Electronic, 5:00pm</td>
</tr>
</tbody>
</table>
Supplemental Readings

Articles on C-Tools:


7. **Recommended:** StataCorp, (2007), Stata Survey Data Release 10 Manual, Stata Statistical Software: Release 10, College Station, TX: StataCorp LP.


9. **Recommended:** SPSS, Inc., (2007), SPSS Complex Samples v16.0, Chicago, IL.


Other references


The ASDA web site (www.isr.umich.edu/src/smp/asda) also provides an updated list of recent publications discussing methods for analyzing complex sample survey data.