

# **Analysis of Complex Sample Survey Data**

*Summer Institute in Survey Research Techniques, 2020*

**Instructors: Brady T. West  
Yajuan Si**

**Course:** SurvMeth 614  
**Dates:** July 6 – July 31, 2020  
**Lectures:** Monday, Tuesday, Thursday, 2:00-4:00 pm  
**Lab/Lecture:** Wednesday, Friday, 2:00-4:00pm  
**Zoom link:** <https://umich.zoom.us/j/93689262189>

## **Instructors:**

Brady T. West  
4118 Institute for Social Research  
(734) 223-9793  
[bwest@umich.edu](mailto:bwest@umich.edu)  
<http://www.umich.edu/~bwest>

Yajuan Si  
4014 Institute for Social Research  
(845) 798-2013  
[yajuan@umich.edu](mailto:yajuan@umich.edu)  
<http://www.umich.edu/~yajuan>

## **Grader:**

Ai Rene Ong  
[aireneo@umich.edu](mailto:aireneo@umich.edu)

## **Topics and Course Requirements**

### **Course Description**

Standard courses on statistical analysis assume that survey data arise 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, stratified multistage sample designs, and measurement errors. Most standard statistical procedures in software packages commonly used for data analysis (e.g. SAS<sup>®</sup>, SPSS<sup>®</sup>, Stata<sup>®</sup>, and R) do not allow the analyst to take most of these properties of survey data into account unless specialized survey procedures are used. Failure to do so can have an important impact on the results of all types of analysis, ranging from simple descriptive statistics to estimates of parameters of multivariate models.

This course provides an introduction to specialized software procedures that have been developed for the analysis of complex sample survey data. The course begins by considering the sampling designs of specific surveys: the National Comorbidity Survey (NCS-R), the National Health and Nutrition Examination Surveys (NHANES), and the Health and Retirement Study (HRS). Relevant design features of the NCS-R, NHANES and HRS include survey weights that take into account differences in probability of selection into the sample and differences in

response rates, as well as stratification and clustering in the multistage sampling procedures used in identifying the sampled households and individuals.

The course will then move on to the introduction of variance estimation techniques that have been developed to take into account stratification and cluster sampling that are properties of the multistage sampling designs used in 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 will then turn to software procedures for commonly used analyses, including testing for between-group differences in means and proportions, regression analysis, logistic regression and multilevel modeling. We will also consider the consequences of nonresponse and missing data on survey analysis and methods for dealing with missing data.

Specialized procedures for survey data analysis from the Stata<sup>®</sup> and R software for data management and analysis will be used in conjunction with the Survey Research Center's own IVEware<sup>®</sup> system to develop course examples and exercises; illustrations will also be presented using software procedures from alternative software packages 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. Participants are welcome to use whatever software they would like for course exercises and projects, provided that the software includes specialized procedures enabling the analyses that will be discussed.

## **Textbook and Class Reading**

The textbook for this course will be *Applied Survey Data Analysis, Second Edition* (ASDA, 2017; publisher: Chapman Hall / CRC Press), co-authored by one of the course instructors (Dr. West, along with his colleagues Steven Heeringa and Pat Berglund). 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.

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 Canvas system. Some of the supplemental readings on Canvas will be assigned, and others will be recommended. **Students are required to have finished all assigned readings prior to the lecture or lab for which they have been assigned.**

## **Prerequisites**

This course is taught at an intermediate level, emphasizing both the theory and practice of the analysis of complex sample survey data. The course does not require rigorous training in mathematics; however, proficiency in basic mathematics, including algebra and functions, is essential. Knowledge of calculus and linear algebra is useful but not required for the course. A first course in survey sampling methods and a basic understanding of sampling concepts such as stratification, cluster sampling and weighting is required. Many students enrolled in this course will have also taken SURVMETH 612, Methods of Survey Sampling. Students should also have familiarity with basic statistical concepts, including point estimates, sampling variance,

confidence intervals, p-values, the maximum likelihood estimation method and simple linear and logistic regression models.

## **Format**

All lecture/lab sessions will take place on Zoom (<https://umich.zoom.us/j/93689262189>), an interactive online audio-video system for remote instruction. **Students are required to attend all sessions**, except for cases of emergencies. In the case of an emergency, students need to notify the instructors in advance. All sessions will be recorded, and students can access the recordings after every session has finished on Canvas.

Students should have a separate camera (desktop or laptop built-in cameras typically have inadequate video quality) and a separate headset with a microphone (desktop and laptop built-in microphones and speakers typically have inadequate audio for this kind of an application). Poor quality desktop or laptop equipment can create echoing, degrading the video or audio for all connected participants.

Lectures on Mondays, Tuesdays and Thursdays will cover basic concepts and methods for each topic, and discuss examples and homework exercises. Lecture notes and examples will be presented using PowerPoint, and copies of these materials will also be provided to each student on the Canvas website for the course. Questions are welcomed during lectures, and discussion of the topics is encouraged.

The lecture/computer lab sessions on Wednesdays and Fridays will allow students to analyze actual survey data using the methods learned during the lectures, with assistance from the instructors, and also perform exploratory analyses and generate hypotheses for their final analysis projects. Lab notes with detailed Stata and R syntax will be handed out at the beginning of the course. Students should review the scheduled lab exercise prior to each Wednesday and Friday lab session.

## **Grading**

The course grading will be based on two criteria:

- > Completion of 5 homework assignments (50%)
- > A final class project (50%)

The homework assignments will typically involve carrying out an analysis of a specified survey data set. These analyses can be done on a student's own PC or laptop. Students are encouraged to work in groups on the homework assignments, and students are welcome to use whatever software they would like (Stata, SAS, R, SPSS, etc.), as long as procedures appropriate for the analysis of complex sample survey data are employed. The work that is ultimately submitted must be done by each student. The five homework assignments are due at the beginning of the class session on the specified due dates.

## **Final Class Project**

The primary aims of this course are to provide class participants with instruction in the concepts and practice 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 teams of students to develop a final project paper based on an independent analysis of a survey data set. The projects will be collaborative efforts of teams of 3-5 students, depending on final enrollment counts, and each team will submit a single team project. The survey data set may be identified by each team or chosen from a list of course data sets.

Work on the final project paper will begin in weeks 1 and 2 with a topic search and investigation of potential data sets. Selection of a project survey data set and topic will be finalized in week 2, and basic descriptive analyses and multivariate analyses will be conducted and reviewed by the instructors in weeks 2-4. A preliminary draft of the final paper with the initial sections (background, literature review, data and methods) will be due **Monday, July 27**. Students are expected to finalize the analysis for their project and complete writing their final project paper during weeks 3 and 4 of the course, and time will be allotted for students to complete these tasks during the final week of the course. The final team paper will be due to the instructors in electronic format at 5:00 pm on **Friday, July 31**. The instructors will be available throughout the course to assist students in each successive phase of the development of the final project paper.

## Course Syllabus

### SurvMeth 614: Analysis of Complex Sample Survey Data Summer, 2020

<u>Date</u>	<u>Day</u>	<u>Type</u>	<u>Topic</u>
July 6	Mon	Lecture 1 (West)	Survey estimation and inference for complex designs (Part 1). Complex sample designs, survey estimation and inference. Multi-stage designs, stratification, cluster sampling, weighting, item missing data, finite population corrections.
July 7	Tues	Lecture 2 (West)	Survey estimation and inference for complex designs (Part 2). Construction and evaluation of survey weights. Sampling distributions, confidence intervals. Design effects. Course data sets. <b>Homework #1 distributed. Course Projects Introduced.</b>
July 8	Weds	Lecture 3, Lab 1 (Si)	Sampling error calculation models; ultimate clusters. Preparing for survey data analysis. Lab 1: Introduction to the Stata and R software, and becoming acquainted with the course data sets.
July 9	Thurs	Lecture 4 (West)	Sampling error estimation for descriptive statistics. Taylor Series linearization method. Sampling error calculation models; ultimate clusters. Sampling error estimation for descriptive statistics using specialized software procedures. Software review. <b>Homework #1 Due. Homework #2 distributed.</b>
July 10	Fri	Lab 2 (Si)	Sampling error estimation for descriptive statistics.
July 13	Mon	Lecture 5/6 (West)	Replication Methods for Variance Estimation. Jackknife Repeated Replication (JRR). Balanced Repeated Replication (BRR). Replication methods in Stata. Methods for categorical data.
July 14	Tues	Lecture 6/7 (West)	Methods for categorical data (continued). Estimation and inference for special statistics (percentiles, indices). Subpopulation estimates. Functions of survey estimates including differences and indices. <b>Homework #2 due. Homework #3 distributed. Prospectus describing decision on topic and data set for Final Course Project due.</b>
July 15	Wed	Lecture 8, Lab 3	Lab 3: Sampling errors for subpopulation estimates. Bivariate analysis (cross-tabulation). Hypothesis testing for contrasts of

<u>Date</u>	<u>Day</u>	<u>Type</u>	<u>Topic</u>
		(Si)	subpopulation estimates. Lecture 8: Linear regression analysis review.
<b>July 16</b>	Thurs	Lecture 8/9 (West)	Regression analysis of complex sample survey data.
<b>July 17</b>	Fri	Lab 4 (Si)	Lab 4: Regression analysis computational exercise. <b>Homework #3 due. Homework #4 distributed.</b>
<b>July 20</b>	Mon	Lecture 10 (West)	Logistic regression (Part 1). <b>Introduction and design/methods for course project due.</b>
<b>July 21</b>	Tues	Lecture 11 (West)	Logistic regression (Part 2). Poisson regression. <b>Homework #4 due. Homework #5 distributed.</b>
<b>July 22</b>	Wed	Lab 5 (Si)	Lab 5: Generalized linear models for complex sample survey data. Poisson regression. Simple logistic regression.
<b>July 23</b>	Thurs	Lecture 12 (West)	Multinomial, ordinal logistic regression. Other GLMs. Hypothesis testing.
<b>July 24</b>	Fri	Lab 6 (Si)	Lab 6: Multinomial logistic regression models, continued. Examples of interpreting estimated coefficients, testing hypotheses, and making inferences. <b>Homework #5 due.</b>
<b>July 27</b>	Mon	Lecture 13 (West)	Survival Analysis with Complex Sample Survey Data. <b>Preliminary draft of course project due.</b>
<b>July 28</b>	Tues	Lecture 14 (West)	Imputation of item missing data. Multiple imputation inference for survey data.
<b>July 29</b>	Wed	Lab 7 (Si)	Lab 7: Nonresponse adjustment, imputation
<b>July 30</b>	Thurs	Lecture 15 (West)	Model-based and design-based approaches to survey data analysis. Multilevel models for complex sample survey data.

<u>Date</u>	<u>Day</u>	<u>Type</u>	<u>Topic</u>
July 31	Fri	Lab 8 (Si)	Work session. <b>Final project due, 5pm.</b>

**SurvMeth 614: Analysis of Complex Sample Survey Data  
Summer 2020 Reading Assignments**

\* Assigned Readings need to be completed prior to the onset of the indicated class.

<b>Class Date</b>	<b>Topic</b>	<b>Assigned Readings*</b>
July 6	Survey estimation and inference for complex sample designs (Part 1). Complex sample designs, survey estimation and inference. Multi-stage designs, stratification, clustering, weighting, item missing data, finite population corrections.	<b>BOOK / OTHER:</b> 1. Syllabus 2. Assigned Readings 3. References 4. Chapters 1 and 2, ASDA
July 7	Survey estimation and inference for complex sample designs (Part 2). Models and assumptions for inference from complex sample data. Sampling distribution, confidence intervals. Design effects. Introduction of course data sets.	<b>BOOK:</b> 1. Chapter 3 (3.1 to 3.5), ASDA  <b>CANVAS:</b> 1. Kessler (1994)
July 8 (lab)	Sampling error calculation models; ultimate clusters. Preparing for survey data analysis.	<b>BOOK:</b> 1. Chapter 4, ASDA
July 9	Sampling error estimation for descriptive statistics. Taylor Series linearization method. Sampling error estimation for descriptive statistics using statistical software. Software review.	<b>BOOK:</b> 1. Chapter 3 (3.6.1 to 3.6.2), ASDA 2. Appendix A, ASDA (browse)  <b>CANVAS:</b> 1. Rust (1985) 2. Siller and Tompkins (2005)
July 10 (lab)	Sampling error estimation for descriptive statistics.	<b>BOOK:</b> 1. Chapter 5 (5.1 to 5.3), ASDA  <b>CANVAS:</b> 1. Kreuter and Valliant (2007)
July 13	Replication Methods for Variance Estimation. Jackknife Repeated Replication (JRR). Balanced Repeated Replication (BRR). Replication methods in Stata®.	<b>BOOK:</b> 1. Chapter 3 (3.6.3 to 3.8), ASDA  <b>CANVAS:</b> 1. Kovar et al. (1988)

July 14	Estimation and inference for special statistics (percentiles, indices). Subpopulation estimates. Functions of survey estimates including differences and indices.	<b>BOOK:</b> 1. Chapter 5 (5.3 to 5.6), ASDA 2. Chapter 6, ASDA
July 15 (lab)	Sampling errors for subpopulation estimates. Bivariate analysis (cross-tabulation). Hypothesis testing for contrasts of subpopulation estimates.	<b>None (catch up on previous assigned readings).</b>
July 16	Linear Regression (Part 1, Part 2).	<b>BOOK:</b> 1. Chapter 7, ASDA
July 17 (lab)	Linear Regression Analysis Computational Exercise.	<b>None (catch up on previous assigned readings).</b>
July 20	Logistic Regression (Part 1).	<b>BOOK:</b> 1. Chapter 8 (through 8.5), ASDA  <b>CANVAS:</b> 1. Hosmer and Lemeshow (2000)
July 21	Logistic Regression (Part 2). Poisson Regression.	<b>BOOK:</b> 1. Chapter 8 (through end), ASDA 2. Chapter 9 (9.4), ASDA  <b>CANVAS:</b> 1. Archer and Lemeshow (2006)
July 22 (lab)	GLM Analysis Methods for Complex Sample Survey Data.	<b>None</b>
July 23	Multinomial, ordinal logistic regression. Other GLMs. Hypothesis testing.	<b>BOOK:</b> 1. Chapter 9 (9.1 to 9.3), ASDA
July 24 (lab)	Multinomial and ordinal logistic regression models. Examples of interpreting estimated coefficients, testing hypotheses, and making inferences.	<b>None (catch up on previous assigned readings).</b>

July 27	Survival analysis and event history analysis.	<b>BOOK:</b> 1. Chapter 10, ASDA
July 28	Imputation of item missing data. Multiple imputation inference for survey data.	<b>BOOK:</b> 1. Chapter 12 (ASDA) <b>CANVAS:</b> 1. Raghunathan et al. (2001)
July 29 (lab)	Multiple Imputation Analysis.	<b>CANVAS:</b> 1. Carlin et al. (2008)
July 30	Multilevel models for complex sample survey data.	<b>BOOK:</b> 1. Chapter 13 (ASDA) <b>CANVAS:</b> 1. Rabe-Hesketh and Skrondal (2006)
July 31 (lab)	Work Day.	<b>None.</b>

## Required and Recommended Readings:

1. **Recommended:** Heeringa, S., and Liu, J. (1997), Complex sample design effects and inference for mental health survey data, *International Journal of Methods in Psychiatric Research*, 7, Whurr Publishers Ltd. – Pages 221 – 230.
2. **Required:** Kessler, R. (1994), The National Comorbidity Survey of the United States, *International Review of Psychiatry*, 6. – Pages 365 – 376.
3. **Recommended:** Little, R.J.A., Lewitzky, S., Heeringa, S., Lepkowski, J., and Kessler, R.C. (1997), Assessment of Weighting Methodology for the National Comorbidity Survey, *American Journal of Epidemiology*, 146 (5). – Pages 439 – 449.
4. **Required:** Rust, K. (1985), Variance Estimation for Complex Estimators in Sample Surveys, *Journal of Official Statistics*, 1 (4), Statistics Sweden Publishing Service. Pages 381 -397.
5. **Required:** Siller, A. B., and Tompkins, L. (2005), The Big Four: Analyzing Complex Sample Survey Data Using SAS, SPSS, STATA, and SUDAAN, SUGI 31, Paper 172-31.
6. **Required:** Kreuter, F., and Valliant, R. (2007), A survey on survey statistics: What is done and can be done in Stata, *The Stata Journal*, Volume 7, Number 1, pages 1 – 21.
7. **Required:** Kovar, J.G., Rao, J.N.K., and Wu, C.F.J. (1988), Bootstraps and other methods to measure errors in survey estimates, *The Canadian Journal of Statistics*, 16. – Pages 25 – 45.
8. **Recommended:** Binder, D.A., Gratton, M., Hidiroglou, M.A., Kumar, S., and Rao, J.N.K. (1984), Analysis of Categorical Data from Surveys with Complex Designs: Some Canadian Experiences, *Survey Methodology*, 10 (2), 1989. – Pages 141 – 156.
9. **Recommended:** Kish, L. and Frankel, M. (1974), Inference from Complex Samples, *The Journal of the Royal Statistical Society Series B (Methodological)*, 36 (1). Pages 1 – 37.
10. **Recommended:** West, B.T., Berglund, P., and Heeringa, S.G. (2008), A Closer Examination of Subpopulation Analysis of Complex Sample Survey Data, *The Stata Journal*, 8(4), 520-531.
11. **Recommended:** Kott, P. (1991), A Model-Based Look at Linear Regression With Survey Data, *The American Statistician*, 45 (2). –Pages 107 -112.
12. **Recommended:** Binder, D. (1983), On the Variances of Asymptotically Normal Estimators from Complex Surveys, *International Statistical Review*, 51. –Pages 279 -292.
13. **Required:** Hosmer, D.W. and Lemeshow, S. (2000), Application of Logistic Regression with Different Sampling Models, *Applied Logistic Regression*, Second edition. – Chapter 6, pages 203 – 222.

14. **Required:** Archer, K.J., and Lemeshow, S. (2006), Goodness-of-fit test for a logistic regression model fitted using survey sample data, *The Stata Journal*, 6(1), 97-105.
15. **Recommended:** Kalton, G. and Kasprzyk, D. (1986), The Treatment of Missing Survey Data, *Survey Methodology*, 12 (1), Statistics Canada. –Pages 1 -16.
16. **Recommended:** Fay, R. (1996), Alternative Paradigms for the Analysis of Imputed Survey Data, *Journal of the American Statistical Association*, 91 (434). –Pages 490 – 498.
17. **Required:** Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, 27 (1). –Pages 85-95.
18. **Required:** Carlin, J.B., Galati, J.C., and Royston, P. (2008), A New Framework for Managing and Analyzing Multiply Imputed Data in Stata, *The Stata Journal*, 8(1), 49-67.
19. **Recommended:** Snijders, T. (2001), Sampling, *Multilevel Modeling of Health Statistics*, New York: Wiley. –Pages 159 – 174.
20. **Recommended:** Hansen, M., Madow, W., and Tepping, B. (1983), An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys, *Journal of the American Statistical Association*, 78 (384). -Pages 776 – 793.
21. **Recommended:** Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998), Weighting for Unequal Selection Probabilities in Multilevel Models, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 60 (1), -Pages 23-40.
22. **Required:** Rabe-Hesketh, S., and Skrondal, A. (2006), Multilevel modeling of complex survey data, *Journal of the Royal Statistical Society, Series A*, 169 (4), 805-827.