Analysis Methods for Complex Sample Survey Data (SURVMETH 614)

Summer Institute in Survey Research Techniques, 2018

Instructors:

Brady T. West
4118 ISR / 3550 Rackham
(734) 647-4615
bwest@umich.edu
www.umich.edu/~bwest

Yajuan Si
4014 ISR
(734) 764-6935
yajuan@umich.edu
www.umich.edu/~yajuan

Course:	**SURVMETH 614**
Dates:	**June 4 - July 27, 2018**
Lectures:	**Monday, Wednesday, 1:00 - 3:00 pm**
Computer Lab:	**Friday, 1:00 - 3:00 pm**
Locations:	**Lectures – 4128 LSA Building**
	Computer Labs – 2302 School of Education (SoE) Building

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 unequal probabilities of selection, stratified multistage sample designs, and missing data. Most standard statistical procedures in software packages commonly used for data analysis (e.g. R, SAS®, SPSS®, and Stata®) do not allow the analyst to take 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 analyses, ranging from simple descriptive statistics to analytic 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 three specific surveys: the National Comorbidity Survey-Replication (NCS-R), the National Health and Nutrition Examination Surveys (NHANES), and the Health and Retirement Survey (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 response rates, as well as the stratification and clustering in multistage sampling procedures used to identify the sampled households and individuals. Example designs from other surveys (e.g., the European Social Survey, or ESS) will be considered as well.

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, which are properties of the multistage sampling designs used by most major survey programs. These techniques 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 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 SAS® and Stata® systems for data management and analysis will be used in conjunction with the Survey Research Center's own IVEware® system to develop course examples and exercises; illustrations will also be presented using procedures from the R and SUDAAN® software packages that have been specifically designed for the analysis of survey data. Data from the NCS-R, NHANES, HRS, and ESS will be used to illustrate the various analysis procedures covered during the course in lab sessions. Homework exercises will enable students to practice analyses using the same three data sets used in the lab sessions.

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 Ulrich's book store (https://www.bkstr.com/ulrichsstore/home), or 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 surveys consider purchasing a copy of *Analysis of Complex Sample Survey Data*, authored by Skinner, Holt and Smith (1989). This text is out of print but it may be possible to locate a copy on Amazon.com or through other book reselling services.

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

Lectures on Mondays and Wednesdays will cover basic theory and methods for each topic, and discuss examples and homework exercises. Students are encouraged to ask questions about the lecture material or the assigned readings, and discuss the topics with other students. Lecture notes and examples will be presented using PowerPoint, and electronic copies of these materials will also be provided to each student on the course Canvas website, along with other relevant information and web links pertinent to the course.

The computer lab sessions on 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 class projects. Lab notes with detailed syntax for the computer packages will be made available on the Canvas site at the beginning of the course. **Students should review the scheduled lab exercise prior to the Friday lab sessions.**

Grading

The course grading will be based on two criteria:

1. Completion of 5 homework assignments (50%)
2. A final class project (50%)

The homework assignments will typically involve carrying out an analysis of a specified survey data set (see Course Description above) and providing interpretation of results. These analyses can be done on a student's own PC or laptop, or on workstations that are available in the University's on-campus computing facilities. The software packages that will be used include R, SAS®, Stata®, and SRC's IVEware®. Students are encouraged to work in groups on the homework assignments. However, **the work that is submitted must be completed individually 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 theory 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 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 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 3, and basic descriptive analyses and multivariate analyses will be conducted and reviewed by the instructors in weeks 4-6. A preliminary draft of the final paper with the initial sections (background, literature review, data and methods) will be due **Friday, July 13**. Students are expected to finalize the analysis for their project and complete writing their final project paper during weeks 7 and 8 of the course, and lab time will be allotted for students to complete these tasks during the final week of the course. The final paper will be due to the instructors in electronic format at 5:00 pm on **Friday, July 27**. The instructors will be available throughout the course to assist students in each successive phase of the development of the final project paper.

Attendance

Students are expected to attend each lecture, attend each lab session, participate in class discussions, and complete all homework and lab assignments on time. If you must miss a lecture or lab, please let the Drs. West/Si know in advance.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please contact Services for Students with Disabilities (SSD) office to help us determine appropriate academic accommodations. SSD (734-763-3000; http://ssd.umich.edu) typically recommends accommodations through a Verified Individualized Services

and Accommodations (VISA) form. Any information you provide is private and confidential and will be treated as such.

Academic Conduct

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at the Rackham web site for the University of Michigan: http://www.rackham.umich.edu/policies/academic_policies/section10/.

Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

Course Syllabus

SURVMETH 614: Analysis Methods for Complex Sample Survey Data
Summer, 2018

Date	Day	Type	Topics / Assignments
June 4	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.
June 6	Wed	Lecture 2 (West)	**Survey estimation and inference for complex designs (Part 2).** Models and assumptions for inference from complex sample survey data. Sampling distributions, confidence intervals. Design effects. Introduction of course data sets. **Homework #1 distributed. Course Projects Introduced.**
June 8	Fri	Lecture 3 / Lab 1 (Si and West)	**Sampling error calculation models; ultimate clusters.** Preparing for survey data analysis. **Lab 1:** Introduction to the course computing facilities, and introduction to the NCS-R, HRS and NHANES data sets.
June 11	Mon	Lecture 4 (West)	**Sampling error estimation for descriptive statistics.** The Taylor Series Linearization (TSL) method. Sampling error estimation for descriptive statistics using statistical software. Software review.
June 13	Wed	Lecture 5 (West)	**Replication Methods for Variance Estimation.** Jackknife Repeated Replication (JRR). Balanced Repeated Replication (BRR). Bootstrapping. Replication methods in Stata® and IVEware®.
June 15	Fri	Lab 2 (Si and West)	**Lab 2:** Sampling error estimation for descriptive statistics. **Homework #1 Due. Homework #2 distributed.**
June 18	Mon	Lecture 6 (West)	**Estimation and inference for special statistics (percentiles, indices).** Subpopulation estimates. Functions of survey estimates including differences and indices.
June 20	Wed	Lecture 7 (Si)	**Methods for Categorical Data.** Rao-Scott and related hypothesis tests. Odds ratios and relative risks.
June 22	Fri	Lab 3 (Si and West)	**Lab 3:** Sampling errors for subpopulation estimates. Bivariate analysis (cross-tabulation). Hypothesis testing for contrasts of subpopulation estimates. (continued on next page…)

Date	Day	Type	Topics / Assignments
			Homework #2 due. Homework #3 distributed. Prospectus describing topic and data set for Final Course Project due.
June 25	Mon	Lecture 8 (Si)	**Linear Regression (Part 1).**
June 27	Wed	Lecture 9 (West)	**Linear Regression (Part 2).** **Homework #3 due. Homework #4 distributed.**
June 29	Fri	Lab 4 (Berglund)	**Lab 4:** Linear Regression Analysis Computational Exercise.
July 2	Mon	Lecture 10 (West)	**Logistic Regression.**
July 4	Wed	**July 4th Holiday**	**No class!**
July 6	Fri	Lab 5 (Si)	**Lab 5:** Logistic Regression Analysis Computational Exercise. **Homework #4 due. Homework #5 distributed.**
July 9	Mon	Lecture 11 (Si)	**Multinomial, ordinal logistic regression.** Other generalized linear models (GLMs). Hypothesis testing.
July 11	Wed	Lecture 12 (West)	**Poisson and negative binomial regression.**
July 13	Fri	Lab 6 (Si)	**Lab 6:** Multinomial and ordinal logistic regression models. Examples of interpreting estimated coefficients, testing hypotheses, and making inferences. **Homework #5 due. Introduction and design/methods for course project due.**
July 16	Mon	Lecture 13 (West)	**Survival analysis and event history analysis.**
July 18	Wed	Lecture 14 (West)	**Imputation of item missing data.** Multiple imputation inference for survey data.
July 20	Fri	Lab 7 (Si and West)	**Lab 7:** Multiple Imputation Computational Exercise.

Date	**Day**	**Type**	**Topics / Assignments**
July 23	Mon	Lecture 15 (West)	**Multi-level models for complex sample survey data.**
July 25	Wed	Work Day	**Open lab for work on final projects.**
July 27	Fri	Work Day	**Open lab for work on final projects.**
			Final project due, 5pm.

SurvMeth 614: Analysis of Complex Sample Survey Data
Summer 2018 Reading Assignments

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

Class Date	Topic	Assigned Readings*
June 4	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.	**BOOK / OTHER:** 1. Syllabus 2. Assigned Readings 3. References 4. Chapters 1 and 2, ASDA
June 6	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.	**BOOK:** 1. Chapter 3 (3.1 to 3.5), ASDA **CANVAS:** 1. Kessler (1994)
June 8 (lab)	Sampling error calculation models; ultimate clusters. Preparing for survey data analysis.	**BOOK:** 1. Chapter 4, ASDA
June 11	Sampling error estimation for descriptive statistics. Taylor Series linearization method. Sampling error estimation for descriptive statistics using statistical software. Software review.	**BOOK:** 1. Chapter 3 (3.6.1 to 3.6.2), ASDA 2. Appendix A, ASDA (browse) **CANVAS:** 1. Rust (1985) 2. Siller and Tompkins (2005)
June 13	Replication Methods for Variance Estimation. Jackknife Repeated Replication (JRR). Balanced Repeated Replication (BRR). Replication methods in Stata® and IVEware®.	**BOOK:** 1. Chapter 3 (3.6.3 to 3.8), ASDA **CANVAS:** 1. Kovar et al. (1988)
June 15 (lab)	Sampling error estimation for descriptive statistics.	**BOOK:** 1. Chapter 5 (5.1 to 5.3), ASDA **CANVAS:** 1. SAS, Stata, SPSS documents (browse) 2. Kreuter and Valliant (2007)

June 18	Estimation and inference for special statistics (percentiles, indices). Subpopulation estimates. Functions of survey estimates including differences and indices.	**BOOK:** 1. Chapter 5 (5.3 to 5.6), ASDA **CANVAS:** 1. West et al. (2008)
June 20	Methods for Categorical Data.	**BOOK:** 1. Chapter 6, ASDA
June 22 (lab)	Sampling errors for subpopulation estimates. Bivariate analysis (cross-tabulation). Hypothesis testing for contrasts of subpopulation estimates.	**None (catch up on previous assigned readings).**
June 25	Linear Regression (Part 1).	**BOOK:** 1. Chapter 7 (through 7.3.3), ASDA
June 27	Linear Regression (Part 2).	**BOOK:** 1. Chapter 7 (through end), ASDA
June 29 (lab)	Linear Regression Analysis Computational Exercise.	**None (catch up on previous assigned readings).**
July 2	Logistic Regression (Part 1).	**BOOK:** 1. Chapter 8 (through 8.5), ASDA **CANVAS:** 1. Hosmer and Lemeshow (2000)
July 6 (lab)	Logistic Regression (Part 2). Logistic Regression Analysis Methods for Complex Sample Survey Data.	**BOOK:** 1. Chapter 8 (through end), ASDA **CANVAS:** 1. Archer and Lemeshow (2006)
July 9	Multinomial, ordinal logistic regression. Other GLMs. Hypothesis testing.	**BOOK:** 1. Chapter 9 (9.1 to 9.3), ASDA
July 11	Poisson Regression. Examples of Poisson Regression Analysis from Lab Notes.	**BOOK:** 1. Chapter 9 (9.4), ASDA
July 13 (lab)	Multinomial and ordinal logistic regression models. Examples of interpreting estimated coefficients, testing hypotheses, and making inferences.	**None (catch up on previous assigned readings).**

July 16	Survival analysis and event history analysis.	**BOOK:** 1. Chapter 10, ASDA
July 18	Imputation of item missing data. Multiple imputation inference for survey data.	**BOOK:** 1. Chapter 12 (ASDA) **CANVAS:** 1. Raghunathan et al. (2001)
July 20 (lab)	Multiple Imputation Analysis.	**CANVAS:** 1. Carlin et al. (2008)
July 23	Multilevel models for complex sample survey data.	**BOOK:** 1. Chapter 13 (ASDA) **CANVAS:** 1. Rabe-Hesketh and Skrondal (2006)
July 25 (lab)	Work Day.	**None.**
July 27 (lab)	Work Day.	**None.**

Selected articles on Canvas (others may be added as the course proceeds):

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. **Recommended:** Skinner, C. J., Holt, D., and Smith, T.M.F., Chapter 2 (pt.), Analysis of Complex Surveys, New York: Wiley. – Pages 24 – 58.

5. **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.

6. **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.

7. **Recommended:** Cassell, D. L., Wait Wait, Don't Tell Me…You're Using the Wrong Proc!, SUGI 31, Paper 193-31.

8. **Required:** SAS Institute, Inc. (2008), Introduction to Survey Sampling and Analysis Procedures (Book Excerpt), SAS/STAT 9.2 User's Guide, Cary, NC: SAS Institute Inc.

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

10. **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.

11. **Required:** SPSS, Inc., (2007), SPSS Complex Samples v16.0, Chicago, IL.

12. **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.

13. **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.

14. **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.

15. **Required:** 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.

16. **Recommended:** Kott, P. (1991), A Model-Based Look at Linear Regression With Survey Data, *The American Statistician, 45 (2).* –Pages 107 -112.

17. **Recommended:** Binder, D. (1983), On the Variances of Asymptotically Normal Estimators from Complex Surveys, *International Statistical Review, 51*. –Pages 279 -292.

18. **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.

19. **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.

20. **Recommended:** Kalton, G. and Kasprzyk, D. (1986), The Treatment of Missing Survey Data, *Survey Methodology, 12 (1),* Statistics Canada. –Pages 1 -16.

21. **Recommended:** Fay, R. (1996), Alternative Paradigms for the Analysis of Imputed Survey Data, *Journal of the American Statistical Association, 91 (434).* –Pages 490 – 498.

22. **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.

23. **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.

24. **Recommended:** Snijders, T. (2001), Sampling, *Multilevel Modeling of Health Statistics*, New York: Wiley. –Pages 159 – 174.

25. **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.

26. **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.

27. **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.