Analysis of Complex Survey Data
SURV702
2 credits/4 ECTS
Summer 2020

Instructor(s)
Benjamin M. Reist, PhD, breist@umd.edu

Short Course Description
Analysis of Complex Sample 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.

Course and Learning Objectives
By the end of the course, students will...
- understand the importance of accounting for the effects of complex sample designs on estimation and inference.
- be able to identify how sample design elements impact estimation and inference.
- be able to estimate sampling error using:
  - direct estimators
  - linearization techniques
  - replication methods
- be able to account for complex sample designs in:
  - descriptive analysis for continuous variables
  - categorical data analysis
  - linear regression
  - logistic regression
- be able to use R statistical software to account for the effects of complex sample designs.

Prerequisites
The prerequisites for SURV702 include one or more graduate courses in statistics covering techniques through OLS and logistic regression, a course in applied sampling methods (e.g. SURV625), 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.

Class Structure and Course Concept:
This is an online course using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video recorded lectures and reading the required literature for each unit and then “attending” mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor. Just like in an on-site course, homework will be assigned and graded and there will be a final project at the end of the course.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching videos, participating in online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb, for each credit offered by a course, students can expect to spend one hour per week on in-class activities and three hours per week on out-of-class activities over the span of a full 8-week term. This is a 2-credit course that runs for 8 weeks. Hence, the total average workload is about 8 hours per week.

Mandatory Weekly Online Meetings:
Wednesday, 9:00 to 10:00 AM EST/3:00 to 4:00 PM CET

Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on https://www.elms.umd.edu/. If video participation via Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. Students are encouraged to respond to each other’s questions. Adding and replying to questions is part of the class participation grade.

Students have the opportunity to use the Zoom meeting room set up for this course to connect with peers outside the scheduled weekly online meetings (e.g., for study
groups). Students are encouraged to post the times that they will be using the room to the course website forum to avoid scheduling conflicts. Students are not required to use Zoom and can of course use other online meeting platforms such as Google Hangout or Skype.

**Grading**
Grading will be based on: three criteria

- Class participation (20%)
- Online Discussion Posts (20%)
- Completion of (8) homework assignments (60%)

Grades will be assigned on the following scale:

<table>
<thead>
<tr>
<th>Grade</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>100 - 97</td>
</tr>
<tr>
<td>A</td>
<td>96 - 93</td>
</tr>
<tr>
<td>A-</td>
<td>92 - 90</td>
</tr>
<tr>
<td>B+</td>
<td>89 - 87</td>
</tr>
<tr>
<td>B</td>
<td>86 - 83</td>
</tr>
<tr>
<td>B-</td>
<td>82 - 80</td>
</tr>
</tbody>
</table>

Etc.

The homework assignments and online discussions are described in more detail below. Dates of when assignment will be due are indicated in the syllabus. Late assignments will not be accepted without prior arrangement with the instructor.

**Homework Assignments**
The course homework assignments will be mostly computational exercises that are provided as a handout. 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 computer using R. The data needed for these assignments will be provided to you via the course website [https://www.elms.umd.edu/](https://www.elms.umd.edu/) Basic familiarity with the use of computers is assumed. Ability to work with R is also assumed, but students can easily learn the fundamentals as a part of the course exercises.

Students are allowed to work in groups on the homework assignments if feasible. **However, the work that is submitted must be done by each student; group submissions are not acceptable, and the submissions will be examined for similarity.** Academic integrity as outlined in the respective graduate student policies, referenced bellow, will be required, and apparent violations of these policies will be dealt with on an individual basis.
Assignments are due every Friday at 11:59 PM EDT on the specified dates, without exception. Email questions about the homework should be directed to breist@umd.edu. Assignments are required to be submitted to https://www.elms.umd.edu/ (PDF preferred), with the file name having the following example structure: S702 HW# firstname lastname.pdf. For homework questions and/or assignments involving statistical software, please include all code used to generate results. Each assignment will have an equal contribution to the overall homework component of the course grade and grades of 0 will be given for any missed assignments. Assignments will be graded electronically and returned the week after their due dates, and the solutions will be discussed in class on that date.

**Online Discussions Forums**

Students are expected to participate in the weekly discussion forums. Students are expected to post at least three (3) questions, comments or, answers to other students about the materials covered in the videos and readings of the week in the forum before the meetings (deadline for posting questions is Tuesday, 5:00 PM EST/11:00 PM CET).

**Technical Equipment Needs**

The learning experience in this course will mainly rely on the online interaction between students and the instructor during the weekly online meetings. Therefore we encourage all students in this course to use a web camera and a headset. Decent quality headsets and web cams are available for less than $20 each. We ask students to refrain from using built-in web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use a wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped. Some employers block certain websites, video technology, or browser plugins. Students should test connections to Zoom and the course website before the first course meeting and address any problems with the relevant authorities or JPSM’s IT department at jpsm-itsupport@umd.edu.

**Long 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., R, 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, in particular R. The course begins by considering the sample designs of existing surveys such as the European Social Survey. Relevant design features 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. We will also consider the consequences of non-sampling error. The R software for data management and analysis will be used to develop course examples and exercises. Data from surveys such as the European Social Survey will be used to illustrate the various analysis procedures covered during the course.

Readings

The textbooks for this course will be:


These books will be referred to as ASDA and CS respectively throughout the rest of the syllabus.

Students can purchase the course texts 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 texts, the instructors have prepared a supplemental readings list that includes several review articles. These
supplemental readings are provided in electronic format at https://www.elms.umd.edu/

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 https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf (University of Maryland) and https://www.uni-mannheim.de/en/research/good-research-practice/ (University of Mannheim).

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.

Accommodations for Students with Disabilities
In order to receive services, students at the University of Maryland must contact the Disability Support Services (DSS) office to register in person for services. Please call the office to set up an appointment to register with a DSS counselor. Contact the DSS office at 301.314.7682; http://www.counseling.umd.edu/DSS/.

Students at the University of Mannheim should contact the Commissioner and Counselor for Disabled Students and Students with Chronic Illnesses at http://www.uni-mannheim.de/studienbueros/english/counselling/disabled_persons_and_persons_with_chronic_illnesses/.

Course Evaluation
In an effort to improve the learning experience for students in our online courses, students will be invited to participate in an online course evaluation at the end of the course (in addition to the standard university evaluation survey). Participation is entirely voluntary and highly appreciated.

Class Schedule
Please note that assignments and dates are subject to change. Information (e.g., articles and assignments) posted to the course website supersedes the information noted here.

Unit 1: Analytical error and the effect of sample design on data analysis

Video lecture: available online Wednesday, May 27, 2020

Online meeting: Wednesday, June 3, 2020, 9:00 AM EDT/3:00 PM CEST

Assignment 1 due Friday, June 5, 2020, 2020, 11:59 PM EDT/Saturday, June 6, 2020, 2020, 05:59 AM CEST

Required readings:

Syllabus

ASDA Chapter 1 and Chapter 2

CS Chapter 1 and Appendix B


Unit 2: Techniques for design-based estimation and inference

Video lecture: available online Wednesday, June 3, 2020

Online meeting: Wednesday, June 10, 2020, 9:00 AM EDT/3:00 PM CEST

Assignment 2 due Friday, June 12, 2020, 2020, 11:59 PM EDT/Saturday, June 13, 2020, 2020, 05:59 AM CEST

Required readings:

ASDA Chapter 3 (3.1-3.7)

CS Chapter 2 (2.1-2.3) and 3


**Unit 3: Estimation and inference for descriptive statistics I: data preparation and univariate analysis**

Video lecture: available online Wednesday, June 10, 2020

Online meeting: Wednesday, June 17, 2020, 9:00 AM EDT/3:00 PM CEST

Assignment 3 due Friday, June 19, 2020, 2020, 11:59 PM EDT/Saturday, June 20, 2020, 2020, 05:59 AM CEST

**Required readings:**
ASDA Chapter 4 and Chapter 5 (5.1-5.3)
CS Chapter 4 (4.1-4.3)

**Unit 4: Estimation and inference for descriptive statistics II: percentiles estimation, bivariate and subpopulation analysis**

Video lecture: available online Wednesday, June 17, 2020

Online meeting: Wednesday, June 24, 2020, 9:00 AM EDT/3:00 PM CEST

Assignment 4 due Friday, June 26, 2020, 2020, 11:59 PM EDT/Saturday, June 27, 2020, 2020, 05:59 AM CEST

**Required readings:**
ASDA Chapter 5 (5.4-5.6)
CS Chapter 2 (2.5) and 4 (4.4-4.6)

BREAK WEEK FROM JUNE 28-JULY 5, 2020

Unit 5: Categorical data analysis for complex survey data

Video lecture: available online Wednesday, June 24, 2020

Online meeting: Wednesday, July 8, 2020, 9:00 AM EST/3:00 PM CET

Assignment 5 due Friday, July 10, 2020, 2020, 11:59 PM EDT/Saturday, July 11, 2020, 05:59 AM CEST

Required readings:
ASDA Chapter 6
CS Chapter 2 (2.4)

Unit 6: Linear and Logistic regression for complex survey data

Video lecture: available online Wednesday, July 8, 2020

Online meeting: Wednesday, July 15, 2020, 9:00 AM EDT/3:00 PM CEST

Assignment 6 due Friday, July 17, 2020, 2020, 11:59 PM EDT/Saturday, July 18, 2020, 05:59 AM CEST

Required readings:
ASDA Chapter 7 and Chapter 8
CS Chapter 5 (5.1-5.2) and 6

Unit 7: Regression diagnostics and model evaluation for complex survey data

Video lecture: available online Wednesday, July 15, 2020

Online meeting: Wednesday, July 22, 2020, 9:00 AM EDT/3:00 PM CEST
Assignment 7 due Friday, July 24, 2020, 2020, 11:59 PM EDT/Saturday, July 25, 2020, 05:59 AM CEST

Required readings:


**Unit 8: To weight or not to weight and the effect of non-sampling error**

Video lecture: available online Wednesday, July 22, 2020

Online meeting: Wednesday, July 29, 2020, 9:00 AM EDT/3:00 PM CEST

Assignment 8 due Friday, July 31, 2020, 2020, 11:59 PM EDT/Saturday, August 1, 2020, 2020, 05:59 AM CEST

Required readings:

ASDA Chapter 3 (3.8)

CS Chapter 5 (5.3)


**Note:** Student access to the course website will be revoked four weeks after the final exam.