Syllabus

SURV 702 Analysis of Complex Survey Data

2 credits/4 ECTS

Stefan Zins, PhD

Video lecture by
Stefan Zins, PhD

September 27 – November 22, 2021

Short Course Description

This course covers the methodology for statistical inference with data from surveys that are not based on a simple random sample from a population.

Statistics of interest covered are: totals, functions of totals (e.g., means or ratios), quantiles, proportions of categorical variables and regression coefficients for linear and generalized linear models. The course is mainly focused on the estimation of sampling variances. Students will learn how complex sampling designs, unit-nonresponse, and survey weights affect the variance of estimators and how this variance can be estimated.

A strong emphasis is given to practical examples where students will learn how to implement the taught methods in the statistical software **R**.

Course Objectives

By the end of the course, students will…

- gain the necessary theoretical knowledge to understand the implication of complex sample designs and non-response for statistical inference.
- be able to estimate sampling error outside the simple random sample context.
- understand how survey weights are constructed and how they work, and thus be able to avoid common misconceptions about them.
- be able to do inference and apply statistical tests for non-linear statistics.
- be able to estimate linear and generalized linear regression models with complex survey data.
- be able to use the statistical software R to apply the covered methodology.

Prerequisites

The prerequisites for this course include one or more graduate courses in statistics covering techniques through OLS and logistic regression, a course in applied sampling methods (e.g., SURV626 Sampling I), 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 the statistical software **R**, 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 (re)introduce students to a finite-population inference framework and will provide a review of some common elements of complex sampling designs. However, students should also have knowledge of some 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. However, we do not venture into the details of sampling algorithms or cover the properties for different unequal probability sampling designs, nor do we address the planning of optimal estimation strategies in detail.

While not a prerequisite, knowledge about Maximum Likelihood Estimation at the level provided in the course SURV706 Generalized Linear Models is recommended.

In addition, students are expected to have an introductory level knowledge of differential calculus and linear algebra to be able to follow the notation of the presented theory.

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 prior to participating in 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 exam 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, you can expect to spend approximately 3h/week on in-class-activities and 9 hours per week on out-of-class activities (preparing for class, readings, assignments, projects, studying for quizzes and exams). Therefore, the workload in all courses will be approximately 12h/week. This is a 2-credit/4-ECTS course that runs for 8 weeks. Please note that the actual workload will depend on your personal knowledge.

Mandatory Weekly Online Meetings

Mondays, 12:00 - 1:00 PM EDT / 6:00 - 7:00 PM CEST, starting September 27, 2021

Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on mannheim.instructure.com. 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. In addition, students are encouraged to post questions about the materials covered in the videos and readings of the week in the forum before the meetings (deadline for posting questions is the Friday 4:00 PM EDT / 10:00 PM CEST before the respective weekly online meeting).

Students have the opportunity to use the BigBlueButton feature in Canvas to connect with peers outside the scheduled weekly online meetings (e.g., for study groups). Students are not required to use Canvas Conferences and can of course use other online meeting platforms such as Google Hangouts, Skype or Microsoft Teams.

Daylight savings time ends in Europe on October 31, 2021 and clocks are turned back 1 hour. Daylight saving time ends in the USA on November 7, 2021. Therefore, look carefully at the times of meetings and deadlines! If in doubt, please consider the CET Time (e.g. in Frankfurt) is the OFFICIAL time for all meetings and deadlines.

Grading

Grading will be based on these criteria:

- Class participation (10%)
- Online Discussion Posts (10%)
- Completion of (2) quizzes (10%)
- Completion of (4) homework assignments (30%)
- Final Exam (40 %)

Grades will be assigned on the following scale:

A+ 100 - 97
A 96 - 93
A- 92 - 90
B+ 89 - 87
B 86 - 83
B- 82 - 80
Etc.

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

The final grade will be communicated under the assignment "Final Grade"; in the Canvas course. Please note that the letter grade written in parentheses in Canvas is the correct final grade. The point-grade displayed alongside the letter grade is irrelevant and can be ignored. The homework assignments and online discussions are described in more detail below. Dates of when assignments will be due are indicated in the syllabus. Late assignments will not be accepted without prior arrangement with the instructor.

Technical Equipment Needs

The learning experience in this course will mainly rely on the online interaction between the students and the instructors 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.

Mannheim Business School would also like to officially inform you that, in order to facilitate your participation in this course, your personal data will be processed by and on systems run by MBS and our subcontractors. You can find detailed information in our privacy policy and information for data subjects here.

Long Course Description

Even if it is not apparent to many daily users of statistics, analyzing survey data can be one of the most challenging tasks in inferential statistics. Survey methodologists and statisticians pointing to the so-called Total Survey Error (TSE) have long acknowledged this. Although there is an awareness of the complex error structure, all too often statistical inference is done with methods that implicitly assume rather simple errors that are straightforward to estimate. The application of simple methodology can be justified by a lack of data required to apply estimators that have desirable properties, such as unbiasedness, consistency, and efficiency. However, there is a risk of oversimplification of statistical inference for the sake of being able to produce an estimate. This increases the hazard of wrong conclusions based on false test decisions.

The course will address a well-studied field of survey statistics: How to estimate and conduct statistical tests under complex sampling design and non-response.

First, the students will be (re)introduced to the finite population inference framework and some common features of complex sampling designs. Students will then learn about estimating the sampling variance of linear and non-linear statistics under complex sampling designs and how it can be approximated for simpler estimations. The variance estimation methods will include direct estimates as well as the computational intensive methods of resampling. Then, we will add survey non-response to the theoretical framework, and learn how non-response can be treated through weighting, and what effect this has on the variances of estimators. Students will then be taught how to do statistical inference for linear statistics such as totals and functions of totals, such as means and ratios, as well as for non-linear and non-smooth statistics such as quantiles, all within the framework of complex sampling designs and survey non-response. Because categorical data is a predominant form of data gathered by surveys in the social sciences, the peculiarities of making inference for proportions are covered in a separate unit. The last two units of the course will then address the estimation of linear and generalized linear models under the introduced framework. A practical focus will be given to how these models are estimated with survey weights, leading to the intersection of the design-based inference used in this course, and the model-based inference where these statistical models originate.

The course will have a strong emphasis on applying learned methods given hands-on examples and practical exercises. The R software for data management and analysis will be used for all implementation and application of the taught methods. In particular, the R *survey* package will be a key tool to conduct most analyses.

Readings

Two textbooks that cover theory and some applications will be used for this course:

- Särndal, C. E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media. (MASS)
- Lohr, Sharon L. Sampling: design and analysis. CRC Press, 2019. (DA)

A third textbook mainly covers applications using **R**:

- Lumley, T. (2010). *Complex Survey: A Guide to Analysis using R.* New York: John Wiley & Sons. (CS)

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

Mandatory:

Särndal, C. E., Swensson, B., & Wretman, J. (2003). Model assisted survey sampling. Springer Science & Business Media. (MASS)
ISBN 9780387406206, 0387406204

Lohr, Sharon L. Sampling: design and analysis. CRC Press, 2019. (DA)
ISBN 9780367273415, 0367273411

Lumley, T. (2010). Complex Survey: A Guide to Analysis using R. New York: John Wiley & Sons. (CS)
ISBN 9781118210932, 111821093X

Särndal, C. E., and Lundström, S. (2005). Estimation in Surveys with Nonresponse. Wiley.
ISBN: 9780470011348, 0470011343

Osier, G. Variance Estimation for Measures Indicators of Poverty and inequality using linearization techniques. Survey Research Methods, 2009.
https://doi.org/10.18148/srm/2009.v3i3.369

Recommended:

Wolter, K. (2007). Introduction to variance estimation. Springer Science & Business Media.
ISBN: 9780387350998, 0387350993

Rust, K. F., and Rao, J. N. K. Variance estimation for complex surveys using replication techniques. Statistical Methods in medical research, 1996, Vol. 5, No. 3.
https://doi.org/10.1177/096228029600500305

J.-C. Deville. Variance estimation for complex statistics and estimators: Linearization and residual techniques. Survey Methodology, 1999, Vol. 25, No. 2.

T. Lumley and A. Scott. Fitting regression models to survey data. Statistical Science, 2017, Vol. 32, No. 2.
DOI:10.1214/16-STS605

E.L. Korn and B.I. Graubard, Confidence Intervals For Proportions With Small Expected Number of Positive Counts Estimated From Survey Data. Survey Methodology, 1998. Vol. 24, no. 2

J. N. K. Rao and A. J. Scott. On chi-squared tests for multiway contigency tables with proportions estimated from survey data. Annals of Statistics, 1984, 12:46-60.
DOI: 10.1214/aos/1176346391

D. Pfeffermann. Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? Survey Methodology, 2011, Vol. 37, No. 2.

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 Accessibility & Disability Service (ADS) office to register in person for services. Please call the office to set up an appointment to register with an ADS counselor. Contact the ADS office at 301.314.7682; https://www.counseling.umd.edu/ads/.

Students at the Mannheim Business School should contact the Commissioner and Counsellor 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. Participation is entirely voluntary and highly appreciated.

Sessions

Week 1: Sampling Theory

At the end of this week, students will:

- Know what expected value and variance of an estimator are
- Know what drives the variance of an estimator
- Be familiar with most common elements of complex sampling design

Video lecture: available September 13, 2021
Online meeting: Monday, September 27, 2021, 12:00 - 1:00 PM EDT / 6:00 - 7:00 PM CEST

Required Readings:
MASS Chapters 2.1 – 2.7
DA Chapters 2.1 – 2.4

Week 2: Variance Estimation Techniques

At the end of this week, students will:

- Know how the sampling variance of linear and non-linear statistics under complex sampling designs can be approximated and estimated.
- Know how design effects can be estimated and used.
- Know how the resampling methods Bootstrap and Jackknife work

Video lecture: September 27, 2021
Online meeting: Monday, October 04, 2021, 12:00 - 1:00 PM EDT / 6:00 - 7:00 PM CEST
Assignment 1: Due Wednesday, October 06, 2021, 5:00 PM EDT / 11:00 PM CEST

Required Readings:
MASS Chapters 2.8, 2.10, 3.7.1, 3.7.2, 4.2, 4.3, 4.4.1, 4.6, 5.5, 5.6
DA Chapter 9.1, 9.3.2, 9.3.3

Week 3: Survey Nonresponse

At the end of this week, students will

- Know how non-response influences variance and expected values of estimators
- Know how to compute survey weights, with different techniques, to address
- survey non-response

Video lecture: October 4, 2021
Online meeting: Monday, October 11, 2021, 12:00 - 1:00 PM EDT / 6:00 - 7:00 PM CEST
Online Quiz 1: Due Wednesday, October 13, 2021, 5:00 PM EDT / 11:00 PM CEST

Required Readings:
DA Chapter 8.1 - 8.5
Särndal, C. E., and Lundström, S. (2005). Estimation in Surveys with Nonresponse. Wiley.
Chapters 5, 6.1 – 6.4., 11.1 - 11.4

Week 4: Inference for Totals, Means and Quantiles I

At the end of this week, students will

- Know how to apply the methods learned in units 2 and 3 (with R) to make inference for common statistics, like totals, means and ratios under complex sampling designs.
- Know how to use calibration weights and how to account for them in variance estimation.

Video lecture: October 11, 2021
Online meeting: Monday, October 18, 2021, 12:00 - 1:00 PM EDT / 6:00 - 7:00 PM CEST
Assignment 2: Due Wednesday, October 20, 2021, 5:00 PM EDT / 11:00 PM CEST

Required Readings:
CS Chapters 2.1.2, 2.2, 3.1, 2.2, 7.1 - 7.4.

Week 5: Inference for Totals, Means, and Quantiles II

At the end of this week, students will

- Know how to estimate the sampling variance of non-smooth functions like quantiles and even measures of inequality like a GINI coefficient.
- Know if and how resampling methods, like Bootstrap and Jackknife can be applied outside the Simple Random Sample setting.
- Know how resampling works in conjunction with calibration weights that compensate for unit non-response.

Video lecture: October 18, 2021
Online meeting: Monday, October 25, 2021, 12:00 - 1:00 PM EDT / 6:00 - 7:00 PM CEST
Online Quiz 2: Due Wednesday, October 27, 2021, 5:00 PM EDT / 11:00 PM CEST

Required Readings:
MASS 5.11
CS Chapter 2.3, 2.4.1
Osier, G. Variance Estimation for Measures Indicators of Poverty and inequality using linearization techniques. Survey Research Methods, 2009.

Recommended reading:
Rust, K. F., and Rao, J. N. K. Variance estimation for complex surveys using replication techniques. Statistical Methods in medical research, 1996, Vol. 5, No. 3.
J.-C. Deville. Variance estimation for complex statistics and estimators: Linearization and residual techniques. Survey Methodology, 1999, Vol. 25, No. 2.

Week 6: Analyzing Categorial Survey Data

At the end of this week, students will

- Know about the effect and use of survey weights in model-based estimation approaches.
- Know about the debate of using survey weights when estimating statistical models, i.e., outside the estimation of “descriptive statistics” like totals and means.

Video lecture: October 25, 2021
Online meeting: Monday, November 01, 2021, **1:00 - 2:00 PM EDT** / 6:00 - 7:00 PM CET
Assignment 3: Due Wednesday, November 03, 2021, **6:00 PM EDT** / 11:00 PM CET

Required Readings:
MASS 5.7 and 5.8
DA Chapter 10
CS 2.4.2, 6.3, 2.5

Recommended reading:
E.L. Korn and B.I. Graubard, Confidence Intervals For Proportions With Small Expected Number of Positive Counts Estimated From Survey Data. Survey Methodology, 1998. Vol. 24, no. 2
J. N. K. Rao and A. J. Scott. On chi-squared tests for multiway contigency tables with proportions estimated from survey data. Annals of Statistics, 1984, 12:46-60.

Week 7: Linear Regression with Survey Data

At the end of this week, students will
- Know about the difference between design- and model- based inference.
- Know how to do inference on regression coefficients without any model assumptions, i.e., in a pure design-based setting.

Video lecture: November 1, 2021
Online meeting: Monday, November 08, 2021, 12:00 - 1:00 PM ET / 6:00 - 7:00 PM CET
Assignment 4: Due Wednesday, November 10, 2021, 6:00 PM ET / 11:00 PM CET

Required Readings:
MASS Chapter 5.10
DA Chapter 11.1 – 11.5
CS Chapter 5.2.1, 5.2.2, 5.2.4, and 5.3

Recommended reading:
D. Pfeffermann. Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? Survey Methodology, 2011, Vol. 37, No. 2.

Week 8: Generalized Linear Regression with Survey Data

At the end of this week, students will Know
- about the effect and use of survey weights in model-based estimation approaches.
- Know about the debate of using survey weights when estimating statistical models, i.e., outside the estimation of “descriptive statistics” like totals and means.

Video lecture: November 8, 2021
Online meeting: Monday, November 15, 2021, 12:00 - 1:00 PM ET / 6:00 - 7:00 PM CET

Required Readings:
DA Chapter 11.7
CS Chapter 6.1 and 6.2

Recommended reading:
T.Lumley and A. Scott. Fitting regression models to survey data. Statistical Science, 2017, Vol.32, No.2.

Final Exam

Available from: November 16, 2021, 12:00 - 1:00 PM ET / 6:00 - 7:00 PM CET
Due on: Due November 22, 2021, 12:00 - 1:00 PM ET / 6:00 - 7:00 PM CET