Fundamentals of Survey and Data Science
SURV 400
3 credits/6 ECTS
Spring 2017

Instructors
Florian Keusch, f.keusch@uni-mannheim.de

Short Course Description
The course introduces the student to a set of principles of survey and data science that are the basis of standard practices in these fields. The course exposes the student to key terminology and concepts of collecting and analyzing data from surveys and other data sources to gain insights and to test hypotheses about the nature of human and social behavior and interaction. It will also present a framework that will allow the student to evaluate the influence of different error sources on the quality of data.

Course and Learning Objectives
By the end of the course, students will…

- be able to apply the key terminology used by survey methodologists and data scientists.
- be able to assess the quality of data from different sources based on a data quality framework.
- be able to select an appropriate data source to answer different types of research questions.
- understand the influence of coverage, sampling, and nonresponse on data quality and know how to deal with deficiencies of the data.
- have a clear understanding of the steps involved in data preparation, data processing, data analysis, and data visualization.
- be able to comply with ethical standards in survey research and data science.

Prerequisites
Students are expected to be familiar with basic statistical concepts, such as mean, standard deviation, variance, and distributions (at the level of an undergraduate course), and have exposure to elements of social science perspectives on human behavior.

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 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, 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 12-week term. This is a 3-credit course that runs for 12 weeks. Hence, the average workload is about 12 hours per week.

Mandatory Weekly Online Meetings

Monday, 18:00-18:50 CET/noon-12:50 PM ET
Convert to your local time zone and be aware that daylight saving time in Germany will begin March 26, 2017.

Mandatory weekly meetings will be held online through Zoom. Log onto https://zoom.us/join and enter the meeting ID 746252996 as well as your name. Please mute your microphone and keep it muted until you are ready to contribute to the discussion (this eliminates feedback and background noise). If video participation via Internet is not possible, students can dial in and join the meetings via telephone. This is only to be used as a back-up option.

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 for the relevant week to the forum for that unit by Monday at 16:00 CET/10:00 AM ET. Students are encouraged to respond to each other's questions. Adding and replying to questions is part of the class participation grade.

Grading

Grading will be based on

- Participation in discussion during the weekly online meetings and contributions to the forum (deadline: Monday, 16:00 CET/10:00 AM ET) demonstrating understanding of the required readings and video lectures (10% of grade)
- Weekly online exercises reviewing specific aspects of the material covered (60% of grade)
- A final open-book online exam (30% of grade)

Weekly online homework assignments are due by 23:59 CET/5:59 PM ET on the Monday following the weekly meeting, at which time the homework will automatically close on the course website. Assignments can be completed at any time during the week leading up to the Monday deadline. Therefore, extensions will be granted sparingly and with penalty.

The final exam is open book and open notes, and students are on the honor system (see Academic Conduct below). Students will have 2 hours to complete the exam online, within the course website. The exam will be open from Tuesday, May 30, 2017 at 00:00 CET/6 PM ET until Monday, June 5, 2017 at 23:59 CET/5:59 PM ET, at which point it will close. All students must submit their exams before midnight June 5, 2017.

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.

Long Course Description

The fields of survey methodology and data science draws on theories and practices developed in several academic disciplines – mathematics, statistics, psychology, sociology, computer science, and economics. To become an accomplished professional in these fields requires a mastery of research literatures as well as experience designing, conducting, and analyzing surveys and data from other sources, such as administrative records, social media, or transactions.

This course introduces the student to a set of principles of survey design and data science that are the basis of standard practices in these fields. The course exposes the student to research literatures that use both observational and experimental methods to test key hypotheses about the nature of human behavior and other factors that affect the quality of data. It will also present important statistical concepts and techniques in sample design, execution, and estimation, as well as models of behavior describing errors in responding to survey questions. Thus, both social science and statistical concepts will be presented.

The course uses the concept of total error as a framework to discuss coverage properties of sampling frames and organic data, alternative sample designs and their impacts on

standard errors of statistics, different modes of data collection and generation, the role of interviewers and respondents in surveys, impacts of nonresponse and missing data on statistics, measurement errors in data, data processing, and data/research ethics.

The course is intended as an introduction to the fields of survey methodology and data science, taught at a graduate level. Lectures and course readings assume that students understand basic statistical concepts (at the level of an undergraduate course) and have exposure to elements of social science perspectives on human behavior. For those lacking such a background, supplementary readings are recommended.

Readings

Primary readings will be from the following two volumes:

> Groves, R.M., Fowler, F.J. Jr., Couper, M.P., Lepkowski, J.M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology, 2nd Edition*. New York: Wiley. (available at local bookstores and online retailers)
>
> Peng, R.D. & Matsui, E. (2015). *The Art of Data Science. A Guide for Anyone Who Works with Data*. Leanpub. (available online at https://leanpub.com/artofdatascience).

Additional required and recommended readings will be made available on the course website: jpsmonline.umd.edu. None of the information in the recommended readings will be included on a homework assignment or the final exam.

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

University of Maryland:
http://www.graduate.umaryland.edu/policies/misconduct.html

University of Mannheim:
http://home.sowi.uni-mannheim.de/english_new/sociology/m_a_in_sociology/Documents/PO%20MA%20Sowi_PolSci+Soz_neu20151209_en.pdf (Examination Regulations, esp. Section 29 – Cheating and other Misconduct) and

https://www.uni-mannheim.de/1/english/research/Good%20Research%20Practice/141119-Satzung%20wiss%20FV%20Senat_en.pdf (Statutes of the University of Mannheim on Procedures for Handling Research Misconduct).

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 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 (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: Introduction – How to do survey research and data science

Online meeting: Mon, February 27, 2017, 18:00-18:50 CET/noon-12:50 PM ET

Assignment 1: due Mon, February 27, 2017, 23:59 CET/5:59 PM ET

Readings:

Groves et al. (2009). Chapters 1.4 and 1.5

Peng & Matsui (2015). Chapters 1-3

Leek, J.T. and Peng, R.D. (2015). What is the question? *Science*, 347, 1314-

1315.

Recommended:

Couper, M. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, 7, 145-156.

Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science and Politics*, 48, 80-83.

Groves, R. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75, 861-871.

Unit 2: Quality of Data

Online meeting: Mon, March 6, 2017, 18:00-18:50 CET/noon-12:50 PM ET

Assignment 2: due Mon, March 6, 2017, 23:59 CET/5:59 PM ET

Readings:

Groves et al. (2009). Chapter 2

Biemer, P. (2010). Total Survey Error. Design, implementation, and evaluation. *Public Opinion Quarterly*, 74, 817-848.

Recommended:

EUROSTAT (2007). Handbook on data quality assessment methods and tools.

Morganstein, D. & Marker, D.A. (2012). Continuous quality improvement in statistical agencies. In L. Lyberg et a. (Eds.), *Survey Measurement and Process Quality*, John Wiley & Sons: Hoboken, NJ, 475-500.

Unit 3: Coverage

Online meeting: Mon, March 13, 2017, 18:00-18:50 CET/1 PM-1:50 PM ET

Assignment 3: due Mon, March 13, 2017, 23:59 CET/6:59 PM ET

Readings:

Groves et al. (2009). Chapter 3

Hargittai, E. (2015). Is bigger always better? Potential biases of Big Data derived from social network sites. *The Annals of the American Academy of Political and Social Science*, 659, 63-76.

Kreuter, F. and Peng, R.D. (2014). Extracting information from Big Data: Issues of measurement, inference and linkage. In J. Lane, V. Stodden, S. Bender, and H. Nissenbaum (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, 257-275.

Unit 4: Modes of Survey Data Collection

Online meeting: Mon, March 20, 2017, 18:00-18:50 CET/1 PM-1:50 PM ET

Assignment 4: due Mon, March 20, 2017, 23:59 CET/6:59 PM ET

Readings:

Groves et al. (2009). Chapters 1.3 and 5

Unit 5: Data Generation from Other Sources

Online meeting: Mon, March 27, 2017, 18:00-18:50 CET/noon-12:50 PM ET

Assignment 5: due Mon, March 27, 2017, 23:59 CET/5:59 PM ET

Readings:

Boyd, D. & Crawford, K. (2012). Critical questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Sociology*, 15, 662-679.

Lazar, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in Big Data analysis. *Science*, 343, 1203-1205.

Wells, C. & Thorson, K. (2017). Combining big data and survey techniques to model effects of political content flows in Facebook. *Social Science Computer Review*, 35, 33-52.

Recommended:

AAPOR (2015). AAPOR Report on Big Data.

Schober, M.F, Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F.G. (2016). Social media analyses for social measurement. Public Opinion Quarterly, 80, 180-211.

Unit 6: Sampling I

Online meeting: Mon, April 3, 2017, 18:00-18:50 CET/noon-12:50 PM ET

Assignment 6: due Mon, April 3, 2017, 23:59 CET/5:59 PM ET

Readings:

Groves et al. (2009). Chapter 4.1-4.6

Peng & Matsui (2015). Chapter 6

Recommended:

Battaglia M.P. et al. (2016). Sampling, data collection, and weighting procedures for address-based sample surveys. *Journal of Survey Statistics and Methodology*, 4, 476-500.

Unit 7: Sampling II

Online meeting: Mon, April 10, 2017, 18:00-18:50 CET/noon-12:50 PM ET

Assignment 7: due Mon, April 10, 2017, 23:59 CET/5:59 PM ET

Readings:

Baker, R. et al. (2010). AAPOR report on online panels. *Public Opinion Quarterly*, 74, 711-781.

Recommended:

Baker, R. et al. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

****No Meeting on Mon, April 17, 2017 (Easter Monday)****

Unit 8: Questionnaires and Interviewing

Online meeting: Mon, April 24, 2017, 18:00-18:50 CET/noon-12:50 PM ET

Assignment 8: due Mon, April 24, 2017, 23:59 CET/5:59 PM ET

Readings:

Groves et al. (2009). Chapters 8.1-8.8 and 9.1-9.7

Recommended:

West, B.T. & Blom, A. (2016). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*. First published online November 1, 2016. doi:10.1093/jssam/smw024.

****No Meeting on Mon, May 1, 2017 (International Workers' Day)****

Unit 9: Nonresponse

Online meeting: Mon, May 8, 2017, 18:00-18:50 CET/noon-12:50 PM ET

Assignment 9: due Mon, May 8, 2017, 23:59 CET/5:59 PM ET

Readings:

Groves et al. (2009). Chapters 6 and 10.6

Recommended:

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329-353.

AAPOR (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 9th edition.

Unit 10: Data Preparation, Data Processing, and Data Base Management

Online meeting: Mon, May 15, 2017, 18:00-18:50 CET/noon-12:50 PM ET

Assignment 10: due Mon, May 15, 2017, 23:59 CET/5:59 PM ET

Readings:

Groves et al. (2009). Chapters 10.1-10.5

Foster, I. & Heus, P. (2016). Databases. In I. Foster, R. Ghani, R.S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big Data and Social Science: A Practical Guide to Methods and Tools*. Chapman & Hall, 93-124.

Unit 11: Data Analysis and Data Visualization

Online meeting: Mon, May 22, 2017, 18:00-18:50 CET/noon-12:50 PM ET

Assignment 11: due Mon, May 22, 2017, 23:59 CET/5:59 PM ET

Readings:

Peng & Matsui (2015). Chapters 4, 5, 7, 8, and 10

Yalcin, A. & Plaisant, C. (2016). Information Visualization. In I. Foster, R. Ghani, R.S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big Data and Social Science: A Practical Guide to Methods and Tools*. Chapman & Hall, 243-263.

Unit 12: Survey and Research Ethics

Online meeting: Mon, May 29, 2017, 18:00-18:50 CET/noon-12:50 PM ET

Assignment 12: due Mon, May 29, 2017, 23:59 CET/5:59 PM ET

Readings:

Groves et al. (2009). Chapter 11

Barocas, S. & Nissenbaum, H. (2014). Big Data's end run around anonymity and consent. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, 44-75.

Final Exam

Online due: Mon, Jun 5, 2017, 23:59 CET/5:59 PM ET