JOINT PROGRAM IN SURVEY METHODOLOGY

Synthetic Data: Balancing Confidentiality and Quality in Public Use Files

Monday, Feb 24 - Friday, March 07, 2025
An Online short course sponsored by the Joint Program in Survey Methodology

JÖRG DRECHSLER
Distinguished Researcher, Institute for Employment Research, Germany

JERRY REITER
Professor of Statistical Science, Duke University

COURSE OBJECTIVES

This short course will provide a detailed overview of the topic, covering all important aspects relevant for the synthetic data approach. Starting with a short introduction to data confidentiality in general and synthetic data in particular, the short course will discuss the different approaches to generating synthetic datasets in detail. Possible modeling strategies and analytical validity evaluations will be assessed and potential approaches to quantify the remaining risk of disclosure will be presented. The course will also briefly describe how synthetic data could be used with differential privacy. All steps will be illustrated using simulated and real data examples in R.

WHO SHOULD ATTEND

The course intends to summarize the state of the art in synthetic data. The main focus will be on practical implementation and not so much on the motivation of the underlying statistical theory. Participants may be academic researchers or practitioners from statistical agencies working in the area of data confidentiality and data access. Some background in Bayesian statistics and R is helpful but not obligatory.

JOINT PROGRAM IN SURVEY METHODOLOGY

INSTRUCTORS

JÖRG DRECHSLER Jörg is distinguished researcher at the Department for Statistical Methods at the Institute for Employment Research in Nürnberg, Germany. He received his PhD in Social Science from the University in Bamberg in 2009 and his Habilitation in Statistics from the Ludwig-Maximilians-Universität in Munich in 2015. He is also an Associate Research Professor in the Joint Program in Survey Methodology at the University of Maryland and Honorary Professor at the University of Mannheim, Germany. His main research interests are data confidentiality and nonresponse in surveys.

JERRY REITER is Professor of Statistical Science at Duke University in Durham, NC. He received his PhD in statistics from Harvard University in 1999. He has developed much of the theory and methodology for synthetic data, as well as supervised the creation of the Synthetic Longitudinal Database. He is the recipient of the 2014 Gertrude M. Cox Award.

CLASS STRUCTURE

The course will be in an online format from Feb 24 to March 07, 2025. Participants will have online access to the course packet (slides) and to the recorded lectures. The recorded lectures will be divided into eight sections (four sections per week) of roughly one hour each. Participants can watch the videos at their own pace, with a recommended viewing of one session per day for each of the four days (Monday – Thursday). **Live group online discussions are scheduled for Friday 02/28/25 and 03/07/25 from 10:30 am to 12 pm.**

These sessions will be used to discuss questions that came up over the week when watching the videos. Students will have the opportunity to submit their questions in advance. A small assignment in R will be posted each week to provide the participants with hands on experience. While these assignments are not mandatory and will not be collected, students are encouraged to work through the assignments prior to the online meetings. Time permitting, solutions to the assignments will be discussed during the online meetings or disseminated to the students after the course.