Synthetic Data: Balancing Confidentiality and Quality in Public Use Files
www.jpsm.org/shortcourses

December 3-4, 2019
Presented at RTI, Washington DC.

JÖRG DRECHSLER
Senior 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 workshop 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 the how synthetic data could be used with differential privacy. To provide the participants with hands on experience, most of the second day will be devoted to practical sessions using R in which the students generate and evaluate synthetic data for various datasets.

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.

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 adjunct associate 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.

COMPUTER
Students should bring their own laptop with R installed. Prior to the course, students should install the latest version of R, which is available for free at http://www.r-project.org/. Registrants should install the R package synthpop, which is available for free from CRAN at cran.r-project.org.
TENTATIVE SCHEDULE

Tuesday: December 3, 2019

08:00 – 09:00  Registrait Check-in
09:00 – 09:30  Overview of data confidentiality
09:30 – 10:30  Introduction to synthetic data
10:30 – 10:45  Coffee break
10:45 – 12:15  Synthetic data models 1
12:15 – 01:45  Lunch
01:45 – 02:45  Synthetic data models 2
02:45 – 03:15  Utility checks
03:15 – 03:30  Coffee break
03:30 – 04:00  Disclosure risk
04:00 – 04:30  Synthetic Data and Differential Privacy
04:30  Adjourn

Wednesday: December 4, 2019

09:00 – 10:00  Exemplary applications
10:00 – 10:15  Coffee break
10:15 – 11:00  Introduction to synthpop package in R
11:00 – 11:45  Students generate synthetic data in small groups
11:45 – 01:15  Lunch.
01:15 – 02:00  Utility checks
02:00 – 03:00  Disclosure checks
03:00 – 03:15  Coffee break
03:15 – 04:00  Discussion among class
04:00 – 04:30  Wrap up
4:30  Adjourn