SWISS STATISTICAL SOCIETY
www.stat.ch

ETH zürich

UNIVERSITÄT BERN

Swiss Statistics Seminar – Spring 2017

Friday, May 5, 2017

University of Bern

14:15 – 15:15

Sebastian Engelke – EPFL

An entropy-based test for multivariate threshold exceedances

15:30 – 16:30

Matthias Templ – ZHAW

Creating Public-Use Synthetic Data From Complex Surveys

16:45 – 17:45

Damian Kozbur – Uni Zurich

Targeted Undersmoothing

Location:
Institut für Exakte Wissenschaften, Sidlerstrasse 5, Lecture Hall B

See also: www.imsv.unibe.ch/research/talks/swiss_statistics_seminars_live/index_eng.html

SWISS STATISTICAL SOCIETY
www.stat.ch

ETH zürich

UNIVERSITÄT BERN

Swiss Statistics Seminar – Spring 2017

Friday, May 5, 2017

University of Bern

Sebastian Engelke

An entropy-based test for multivariate threshold exceedances

Many effects of climate change seem to be reflected not in the mean temperatures, precipitation or other environmental variables, but rather in the frequency and severity of the extreme events in the distributional tails. Detecting such changes requires a statistical methodology that efficiently uses the largest observations in the sample. We propose a simple, non-parametric test that decides whether two multivariate distributions exhibit the same tail behavior. The test is based on the entropy, namely Kullback-Leibler divergence, between exceedances over a high threshold of the two multivariate random vectors. We show that such a type of divergence is closely related to the divergence between Bernoulli random variables. We study the properties of the test and further explore its effectiveness for finite sample sizes. As an application we apply the method to precipitation data where we test whether the marginal tails and/or the extremal dependence structure have changed over time.

Matthias Templ

Creating Public-Use Synthetic Data From Complex Surveys

The production of synthetic datasets has been proposed as a statistical disclosure control solution to generate public use files from confidential data. This is also a tool to create "augmented datasets" to serve as input for micro-simulation models, and the synthetic data sets can be used for design-based simulation studies in general. The performance and acceptability of such a tool relies heavily on the quality of the synthetic data, i.e., on the statistical similarity between the synthetic and the true population of interest. Multiple approaches and tools have been developed to generate synthetic data. They belong to three main groups: synthetic reconstruction, combinatorial optimization, and model-based generation. In addition, methods have been formulated to evaluate the quality of synthetic data. In this presentation, the methods are introduced in an applied and generally understandable fashion. We focus on new concepts for the model-based generation of synthetic data that avoids disclosure problems. In the end, we introduce an open source data synthesizer: simPop is a user-friendly R-package based on a modular object-oriented concept. It provides a highly optimized S4 class implementation of various methods, including calibration by iterative proportional fitting/updating and simulated annealing, and modeling or data fusion by logistic regression, regression tree methods etc. Utility functions to deal with (age) heaping are implemented as well. An example is shown using real data from Official Statistics. The simulated data then serves as input for agent-based simulation and/or microsimulation or can be used as open data for research and teaching.

Damian Kozbur

Targeted Undersmoothing

This talk describes a post-model selection inference procedure, called 'targeted undersmoothing', designed to construct confidence sets for a broad class of functionals of high-dimensional statistical models. These include dense functionals, which may potentially depend on all elements of an unknown high-dimensional parameter. The proposed confidence sets are based on an initially selected model and two additionally selected models, an upper model and a lower model, which enlarge the initially selected model. The procedure is illustrated with two examples. The first example studies heterogeneous treatment effects in a direct mail marketing campaign, and the second example studies treatment effects of the Job Training Partnership Act of 1982.