Kolloquium „Statistische Methoden in der empirischen Forschung“

Wann: 14. Februar 2017, 17:00 – 18:30 Uhr

Wo: Robert Koch-Institut | Nordufer 20 | 13353 Berlin (Wedding),
S41, S42, U9 Westhafen | U9, Bus 142 Amrumer Str

Hans-J. Lenz (FU Berlin)

Aufdeckung von Datentricksereien

Data Fraud is a criminal activity done by at least one person who intentionally acts secretly, out of law or without any agreement to deprive other people of something of value for their own benefit, i.e. profit or prestige. Mostly data fraud is performed by ‘Paste and Copy’. The digital revolution simplifies the work of tricksters.

Data Fraud happened and still happens everywhere in all centuries and in all fields of human activities: Business, economics, politics, science, health care, religious communities, daily life etc.

Data Fraud is extensionally characterized by four fields: *Data Theft, Plagiarism, Manipulation* and *Fabrication*.

Data Theft and spying is by no means limited to the military or secret service area, cf. NSA activities, but is common to industry as industrial spying focusing on top secret data like longterm strategic planning, industrial designs or manufacturing skills. The Internet simplifies thieving the identities of innocent users by criminal gangs. Today, the automobile industry makes us belief that on-board generated data is their own.

Data Plagiarism suppresses referring to the source or provenance of data used by the deceiver. It concerns the cloning of luxurious commercial goods, pictures, music and documents. It became very popular by the current cases of two German secretaries of state, Schavan and von Guttenberg, accused of dissertation plags. Even *Self Plagiarism* is uprising.

Data Manipulation takes existing data and manipulates the content encapsulated in tables, diagrams, figures or (historical) pictures. A famous example is the retouches of a historical picture of Stalin originally with and later without Lew Trotsky. In 2002, the manipulation of measurement errors of field effects by the famous physicist Jan Hendrik Schön caused much hallo.

Finally, *Data Fabrication* generates more or less artificial data by brute force methods for ‘Generating data on demand’ – thus avoiding expensive data recording, time-consuming observations and computing, or running statistically well planned experiments. A dreadful case is given by Diederik Stapel, a Dutch social psychologist, who fabricated experimental data for ‘proving’ doubtful hypotheses until 2011.

There is and will be no omnibus test available to detect data fraud of all kind. However, a bundle of techniques like substring matching and citation based analytics, probability distribution analysis methods, Benford’s Law application, inliers and outlier as well as tests of conformity between data and models exist to give hints for (numerical) data fraud.

We present famous cases spanning more than 2000 years of human mankind, and discuss a bundle of useful tests. Furthermore, we hope that in Science independent authorities like research foundations (NSF, DFG, MPS,…) or other organizations like *ResearchGate, Berlin,* will contribute to an increase transparency and verifiability of published (observational or experimental) data. No doubt, we urgently need an “Intl. Scientific Anti Data Fraud Authority” in science. Scientific research studies financed by public funds should be obliged to deposit their data, enabling ‘just in case’ checks similar to activities in dope control. Examining the sampling scheme and the model selection steps is mandatory. Repeatability and reproducibility is necessary to be checked. Furthermore, the final statistical model, the parameter estimates and the hypothesis tests are of concern.