
# &

# Annual Meeting of the Netherlands Society for Statistics and Operations Research (VVSOR)

*March 20 & 21, 2019*

**LOCATION**

Jaarbeurs Meetup Utrecht

Jaarbeursplein, 3521 AL Utrecht (adjacent to Central Station)

**PROGRAM Wednesday March 20 | Day 1**

12:45 – 13:25 Registration and coffee/tea

13:25 – 13:30 WELCOME by **Fred van Eeuwijk**, president of the VVSOR

13:30 – 14:15 BIG DATA TO IMPROVE POLICY AND DECISION MAKING: THE EXPERIENCE OF STATISTICS NETHERLANDS
**Sofie De Broe**, Center for Big Data Statistics, Heerlen

14:15 – 15:00 WILL DIFFERENTIAL PRIVACY CHANGE THE WAY WE STUDY PEOPLE?
**Daniel Oberski**, Utrecht University

15:00 – 15:30 Break

15:30 – 16:15 THE OBJECTIVITY OF STATISTICS NOW AND IN THE FUTURE
**Sanne Blauw**, numeracy correspondent for *De Correspondent*

16:30 – 17:15 ANNUAL GENERAL MEETING (ALV)

17:15 – 18:30 Snacks and Drinks

19:00 – 23:00 DINNER AND PUB QUIZ BY YOUNG STATISTICIANS
At Stadskasteel Oudaen, Oudegracht 99, 3511 AE Utrecht

**PROGRAM Thursday March 21 | Day 2**

10:00 – 10:25 Registration and coffee/tea

10:25 – 10:30 WELCOME by **Fred van Eeuwijk**, president of the VVSOR

10:30 – 11:15 ABOUT BIG DATA AND PRIVACY: LEGAL, ETHICAL AND TECHNICAL CHALLENGES
**Cecile Schut**, Autoriteit Persoonsgegevens

11:15 – 13:30 CEREMONY OF THE WILLEM R. VAN ZWET AWARD AND THE JAN HEMELRIJK AWARD
Prize winners will be presented by the juries, followed by a short presentation by the laureates

12:00 – 13:30 Lunch

13:30 – 14:15 DATA PRIVACY IN OPTIMAL CAPACITY SHARING
**Ilker Birbil**, Erasmus University Rotterdam

14:15 – 15:00 BAYESIAN VARIABLE SELECTION FOR MENDELIAN RANDOMIZATION
**Apostolos Gkatzionis**, MRC Biostatistics Unit, University of Cambridge

15:00 – 15:15 ANNOUNCEMENT OF THE 63RD ISI WORLD STATISTICS CONGRESS 2021 IN THE NETHERLANDS
**Eric Schulte Nordholt,** Chairman of the National Organizing Committee

15:15 – 15:45 Break

15:45 – 16:30 IN SEARCH OF LOST INFINITIES; WHAT IS THE "N" IN BIG DATA?
**Stephen Senn**, consultant statistician

16:30 – 17:30 Snacks and Drinks

# March 20 | Day 1

13:30 – 14:15

## BIG DATA TO IMPROVE POLICY AND DECISION MAKING: THE EXPERIENCE OF STATISTICS NETHERLANDS


### Sofie De Broe

*Center for Big Data Statistics, Heerlen*


Can we improve policy and decision making by using Big Data in official statistics? At the CBS Center for Big Data Statistics (CBDS) we aim to accelerate the introduction of big data sources in official statistics while taking our responsibility in issues of privacy and security. We believe that using Big Data in our statistical products deepens insights in relevant societal questions and improves both timeliness and level of detail. The mission, scope and ecosystem of CBDS will be elaborated, results in terms of beta or experimental statistics and challenges in the transition from experimental to official statistic will be discussed.

Sofie De Broe is head of methodology and scientific director of the Centre for Big Data Statistics at Statistics Netherlands in Heerlen. She has a master's degree in demography (University of Louvain La Neuve, Belgium), with a bachelor in sociology (University of Ghent, Belgium) and she has a PhD in social statistics/reproductive health from the university of Southampton (UK). After working at the university of Southampton as a teaching fellow, Sofie worked for 6 years at the Office for National Statistics in Titchfield on Improving Migration Estimates. In 2012 she moved to Germany where she taught Research Methods at the university of Duisburg-Essen. Since 2015 Sofie works for Statistics Netherlands and mainly focus on Big Data methods, developing experimental statistics using Big Data and implementing them in the statistical process.

14:15 – 15:00

## WILL DIFFERENTIAL PRIVACY CHANGE THE WAY WE STUDY PEOPLE?


### Daniel Oberski

*Utrecht University*


Science should be open and reproducible, and so the pressure is on to publish data, including data on people. But pressure is also on to protect individual research participants – for which it is necessary to inject randomness in the data. I will discuss why this is true, and what it means for researchers who want to analyze data from people. As I will argue, differential privacy will likely force us to change our ways: we will need to account for "privacy error" in our statistics, increase our sample sizes, make more use of preregistration or other self-limitation where possible, and sometimes, completely change our data collection designs.

Daniel Oberski is an associate professor at Utrecht University's Department of Methodology & Statistics. He is a member of the Young Academy of the Royal Netherlands Academy of Arts and Sciences (KNAW) and received a VENI grant from Netherlands Organisation for Scientific Research (NWO) for: "Developing latent variable techniques that open up a treasure trove for social science". His research focuses on data science methodology, in particular the problem of measurement in the social sciences. To draw accurate substantive conclusions, social scientists need to measure human behavior and opinions reliably and validly. Where this ideal is unattainable, the extent of the problem should be known so it can be accounted for in the substantive analysis. His research has contributed to this by:

1. Estimating measurement error in hundreds of survey questions from the European Social Survey and creating a meta-analysis that predicts the extent of such errors;

2. Developing models that correct multivariate social science analyses for the effects of measurement error while retaining accurate measures of uncertainty about the results;
3. Introducing several novel methods to evaluate the fit of latent variable models, the type of model used to attain the two goals above;
4. Collaborating on substantive social science research and implementing his own and others' methods in user-friendly software.

From 2006–2011 Daniel worked for the European Social Survey (ESS). As a member of the group in charge of the evaluation of question quality in the ESS he worked on multitrait-multimethod (MTMM) models.

**15:30 – 16:15**

## THE OBJECTIVITY OF STATISTICS NOW AND IN THE FUTURE


### Sanne Blauw

*De Correspondent*


Statistics appear to be objective, but are they? Sanne Blauw (1986) started asking herself this question when she was doing research in Bolivia for her PhD in econometrics. Reality seemed way too complex to be grasped with numbers. There were many things that counted, but that she couldn't count. Also, she saw how her own point of view affected her research. After her PhD, she decided to become a journalist for the Dutch online platform *De Correspondent* in order to investigate the question: What should be the role of statistics in society? And - with the rapid development of big data and algorithms - what should it be in the future?

Last Fall Sanne published *Het bestverkochte boek ooit* (*met deze titel*) / *The Biggest Bestseller of All Time* (*with this Title*). In the book she argues against putting numbers on a pedestal. On the other hand, she doesn't want to throw them away altogether. She wants to put them back in their place: next to words. From GDP-numbers to IQ, from nutrition research to international rankings - Sanne makes clear what such numbers do and do not say.

Sanne Blauw (1986) has an MSc in Econometrics (cum laude) and completed her PhD in 2014 with the dissertation 'Well-to-do or Doing Well', on income inequality, trust and happiness. But, she asked herself, can you measure happiness? This question launched her career in journalism. She became Numeracy Correspondent for *De Correspondent*, with one goal: to unveil the bizarre influence of numbers on our lives.

# March 21 | Day 2

**10:30 – 11:15**

## ABOUT BIG DATA AND PRIVACY: LEGAL, ETHICAL AND TECHNICAL CHALLENGES


### Cecile Schut

*Autoriteit Persoonsgegevens*


A lot of big data sources contain personal data: any information relating to an identified or identifiable natural person. So it is clear that the use of big data might have implications for the right to privacy. In Europe, the General Data Protection Regulation demands fair, accurate and non-discrimatory use of personal data, whether it is big data, small data or ordinary data. Is it possible to embed privacy and data protection into big data analysis? Challenges will be illuminated from three different perspectives: the legal, the ethical and the technical viewpoint.

Cecile Schut has a masters' degree in applied mathematics and in public administration. After a few years at KPN Research, she joined Statistics Netherlands in 1997, where she held various positions. In her position as Director of the policy staff at Statistics Netherlands (2013–2017) she was, among other things, responsible for the privacy and quality policy of Statistics Netherlands, including the implementation of the GDPR. From January 2018, she joined the Dutch Data Protection Authority, the independent supervisory body in the Netherlands that fosters and monitors the protection of personal data. As Director of System Supervision, security and technology, Cecile is responsible for offering guidance to organisations and their data protection officers, the assessment of requests for prior consultations, codes of conduct and other GDPR instruments that stimulate organisations to become privacy-proof. Besides, her unit develops high-quality and up-to-date knowledge in the field of security and technology which is necessary for the different supervisory tasks of the Dutch DPA.

**13:30 – 14:15**

## DATA PRIVACY IN OPTIMAL CAPACITY SHARING


### Ilker Birbil

*Erasmus University Rotterdam*


Capacity sharing is arguably one of the best approaches to obtain sustainable and cost-effective use of resources. There exist various mathematical programming tools for optimal resource allocation. However, we still need to convince multiple parties to agree upon sharing their capacities. Even if they give their consent for collaboration, they also rightfully raise their concerns regarding the privacy of their sensitive data used in optimization models. Particularly for resource allocation, linear programming is one of the most frequently used optimization methods in practice. Therefore, in this talk I shall discuss two general ideas to obtain data privacy in linear programming: data masking and problem decomposition. The former idea has also ties with a recently developed research topic known as differential privacy. Along with a presentation of these methodologies, I shall also illustrate their use on application examples from revenue management and logistics. The talk will end with a discussion on some open research questions.

Ilker Birbil has been a faculty member in Erasmus University Rotterdam at the Econometric Institute since 2018. He is serving as an endowed professor for the Chair in Data Science and Optimization. In the past, Ilker worked as a faculty member in Sabancı University, Industrial Engineering Program for 14 years. And a long long time ago, after receiving his B.S. and M.S. degrees in Turkey, he stayed in USA for almost three years for his Ph.D. study. Right after that, he became a post-doctoral research fellow in the Netherlands for two years. Ilker's research interests are parallel and distributed optimization in machine learning,

algorithm development for large-scale optimization problems, data science, revenue management, stochastic dynamic programming. Lately, he is very much interested in data privacy in decision making.

**14:15 – 15:00**

## BAYESIAN VARIABLE SELECTION FOR MENDELIAN RANDOMIZATION


### Apostolos Gkatzionis

*MRC Biostatistics Unit, University of Cambridge*


Mendelian randomization is the use of genetic information to assess the existence of a causal relationship between a risk factor and a disease outcome. It is an application of instrumental variables analysis in the field of statistical genetics, where genetic variants (SNPs) are used as instruments, and has become popular in recent years due to the widespread availability of genetics data.
Recent Mendelian randomization analyses utilize data from large consortia-based Genome-wide Association Studies (GWAS). Individual-level data from such databanks are usually not available due to ethical/privacy considerations, and Mendelian randomization studies rely on summarized data (univariate SNP-trait association estimates and corresponding standard errors) in order to identify genetic instruments strongly associated with the risk factor of interest. The JAM algorithm (Joint Analysis of Marginal summary statistics, Newcombe, Conti and Richardson, 2016) is often used for this task. JAM uses a reversible-jump MCMC procedure to perform SNP selection based on GWAS summary data. It accounts for genetic correlations and can be parallelized to analyse large numbers of SNPs simultaneously.
After providing a brief introduction to Mendelian randomization, we discuss how the JAM algorithm can be used to establish the validity of the instrumental variable assumptions. We propose an extension of the algorithm that augments the JAM posterior with a loss function in order to penalize SNPs having a direct effect on the outcome. The performance of the new algorithm is compared against established Mendelian randomization methods. In a real-data application, we study the effect of blood pressure on the risk of coronary heart disease.

Since January 2017, Apostolos Gkatzionis is a Career Development Fellow (Postdoctoral researcher) at the MRC Biostatistics Unit. He is working with Paul Newcombe and Steve Burgess, and his research is on Mendelian randomization. Before joining the BSU, Apostolos completed a PhD in Statistics at the University of Warwick. His supervisor was Professor David Firth. Briefly, his PhD research was on the analysis and presentation of results of (Bayesian) inference on statistical models containing categorical explanatory variables.
His research interests are Mendelian randomization using genetic variants (SNPs) as instrumental variables to assess the existence of a causal relationship between a biomedical risk factor and a (disease) outcome. It is an approach for causal inference in genetic epidemiology and has become quite popular in recent years.

**15:00 – 15:15**

## THE WORLD STATISTICS CONGRESS 2021

In 2021 the 63rd bi-annual World Statistics Congress of the ISI will take place in The Hague. Eric Schulte Nordholt, chairman of the National Organizing Committee for the WSC2021, will give a short presentation about this major event. The VVSOR is involved in the organization as well.

**15:45 – 16:30**

## IN SEARCH OF LOST INFINITIES; WHAT IS THE "N" IN BIG DATA?

**Stephen Senn**
*Consultant Statistician, Edinburgh*

In designing complex experiments, agricultural scientists, with the help of their statistician collaborators, soon came to realise that variation at different levels had very different consequences for estimating different treatment effects, depending on how the treatments were mapped onto the underlying block structure. This was a key feature of the Rothamsted approach to design and analysis and a strong thread running through the work of Fisher, Yates and Nelder, being expressed in topics such as split-pot designs, recovering inter-block information and fractional factorials. The null block-structure of an experiment is key to this philosophy of design and analysis. However modern techniques for analysing experiments stress models rather than symmetries and this modelling approach requires much greater care in analysis, with the consequence that you can easily make mistakes and often will.

In this talk I shall underline the obvious, but often unintentionally overlooked, fact that understanding variation at the various levels at which it occurs is crucial to analysis. I shall take three examples, an application of John Nelder's theory of general balance to Lord's Paradox, the use of historical data in drug development and a hybrid randomised non-randomised clinical trial, the TARGET study, to show that the data that many, including those promoting a so-called *causal revolution*, assume to be 'big' may actually be rather 'small'. The consequence is that there is a danger that the size of standard errors will be underestimated or even that the appropriate regression coefficients for adjusting for confounding may not be identified correctly.

I conclude that an old but powerful experimental design approach holds important lessons for observational data about limitations in interpretation that mere numbers cannot overcome. Small may be beautiful, after all.

Originally from Switzerland, Stephen Senn was head of the Competence Center for Methodology and Statistics at the Luxembourg Institute of Health (Previously known as CRP-Santé) in Luxembourg, 2011–2018, Professor of Statistics at the University of Glasgow, from 2003 to 2011, and Professor of Pharmaceutical and Health Statistics at University College London from 1995–2003. He has also worked in the Swiss pharmaceutical industry, as a lecturer and senior lecture in Dundee and for the National Health Service in England. He is the author of the monographs *Cross-over Trials in Clinical Research* (1993, 2002), *Statistical Issues in Drug Development* (1997, 2007), *Dicing with Death* (2003) and over 300 scientific publications. In 2001 Stephen Senn was the first recipient of the George C. Challis award for Biostatistics of the University of Florida, in 2008 he gave the Bradford Hill lecture of the London School of Hygiene and Tropical Medicine and in 2009 was awarded the Bradford Hill Medal of the Royal Statistical Society. In 2017 he gave the Fisher Memorial Lecture. He is a Fellow of the Royal Society of Edinburgh and an honorary life member of Statisticians in the Pharmaceutical Industry (PSI) and the International Society for Clinical Biostatistics. He retired in 2018 but is still researching and consulting in statistics.