ABSTRACTS

Modern Statistics and Data Science Education

49th Winter Conference in Statistics

8-12 March 2026

CONTRIBUTED PRESENTATIONS

Teaching Statistics with Simulations

Nicola Orsini

Abstract

Despite the growing importance of statistics in health sciences, teaching methods in higher education tend to follow a standardized approach across disciplines: lectures on theory and methods, followed by practical exercises with predefined datasets. This often presents chal- lenges for students in grasping essential statistical concepts crucial for conducting research. To address these challenges, we propose an engaging learning approach—DICE (Design, Interpret, Compute, Estimate)—aimed at enhancing the statistics learning experience in public health and epidemiology. Through a practical example, we introduce DICE, wherein students collaborate in small groups to plan, generate, analyze, interpret, and communicate their own scientific in- vestigations using simulations. Focused on fundamental statistical concepts such as sampling variability, error probabilities, and statistical model construction, DICE offers a promising ap- proach to integrating substantive knowledge with statistical principles. The materials provided, including computer code, serve as a hands-on tool for both teachers and students.

Coding with challenges: an effective path to data analysis

Linda Hartman

Abstract

The growing need to analyse large and complex data sets means that an increasing number of students, regardless of subject area, need basic knowledge of data analysis. Data literacy – i.e. the ability to handle, visualise and interpret data – is now a key competence for an increasing number of students, and not just in technology and natural sciences.

In this presentation, we share experiences from a relatively new course at Lund University, teaching data analysis and basics of machine learning.

A central part of our approach is inspired by Software Carpentry, where code-along is used as a pedagogical tool. When the teacher codes live and the students follow along step by step, an inclusive and active learning situation is created. This is complemented by well-designed challenges, which give students the opportunity to immediately try out their new skills. The challenges range from small problem-solving tasks to more extensive projects and class competitions, which strengthen both understanding and motivation.

We conclude by discussing our most important lessons learned from the course design and point out possible future development paths for basic data analysis and statistics teaching at Lund University.

Educating Data Scientists in a Modern Statistical Landscape: Experiences from a Nearly Twenty Year Old Master's Programme

Bertil Wegmann

Abstract

The Master's Programme in Statistics and Machine Learning at Linköping University, established nearly twenty years ago, was one of the first programmes of its kind in Sweden. It prepares students to work analytically and critically in today's data driven society, where large and complex information flows are a natural part of modern digital systems. This presentation outlines the programme's structure, its pedagogical integration of statistics, machine learning and programming for advanced data analysis, the diverse backgrounds of its students, and key challenges and strengths within the modern landscape of statistics and data science education. Many of our students find jobs in industry or continue to doctoral studies shortly after completing the programme, highlighting its relevance for both professional and academic futures.

Low-cost teaching activities that support student learning

Maria Karlsson

Abstract

In this presentation, we will showcase examples of successful low-cost teaching activities developed at the Department of Statistics, Umeå University. The activities require limited instructor time yet lead to improvements in student learning. Examples include a workshop designed to increase transparency in the process of assessing (grading) students' exams and a "model text" used to support students in writing data analytics reports. Although the examples originate from specific courses, they can be adopted for most courses in Statistics.

A reflection on the historical development of the subject area of Statistics

Alam Moudud

Abstract

The media outcry and the popularity of the computing technology, and artificial intelligence have pushed the Statistics community to rethink about its longstanding branding. Dedication of the 2024 Nobel prize in Physics and partly in Chemistry, to the contribution in the development of artificial neural network, and artificial intelligence can be considered yet another dictation from the scientific community of the future direction of the filed. The contemporary labour market demands of the computing and soft skills is a non-negligible factor influencing the current trend of the subject. In this review work, the recent development of the subject area of Statistics, particularly the development of Data Science, is presented with reference to the historical milestones in the literature, alongside the related Swedish and European initiatives. It is argued that in this impassionate media outrage it seems the scientific community is undermining (if not missing) the core assignment, as all researchers concentrate too much to the practical applications driven by the contemporary problems, mainly coming from the industry. Using examples from the literature and author's own research, a few statistical core issues regarding the limitations of popular ad hoc (such as cross validation) inferential procedure is exemplified. The examples are brought to highlight the need for core statistical skills in dealing with unconventional data sources in the Data Science era.

POSTERS

Can municipalities mitigate the effects of parental job losses on children's mental health? Valid test when using machine learning methods

Natalia Andreeva

Abstract

This study investigates heterogeneity in the causal effects of parental job displacement on children's mental health within a potential outcomes framework. Using population-wide Swedish register data, we analyze children aged 7–10 whose parents experienced involuntary job loss due to plant closures. Mental health outcomes are measured using administrative records on prescriptions for anxiety and depression medications. We estimate conditional average treatment effects using causal forest algorithms to flexibly model treatment effect heterogeneity in high-dimensional settings. Estimation combines machine learning–based nuisance function learning with doubly robust and orthogonal score functions, ensuring valid statistical inference when using flexible machine learning estimators and accounting for clustering at the municipal level. Our findings indicate that higher municipal spending on elementary schools reduces the negative mental health effects of parental job loss.

Model-Based Functional Clustering with Dependent Errors

Rana Bamdadi, Sara Sjöstedt de Luna, Per Arnqvist, and Natalya Pya Arnqvist

Abstract

In this work, we study and extend the model-based functional clustering framework proposed by Arnqvist and Sjöstedt de Luna (2019). The original model assumes independent measurement errors with constant variance within each functional observation, however, this assumption is often violated in real-world applications due to temporal or serial dependence. We propose a natural extension in which within-curve errors follow a first-order autoregressive (AR(1)) process. To assess the impact of this extension, we conduct extensive simulation studies in which key parameters are systematically varied, including the number of spline basis functions, the AR(1) dependence parameter, and the error variance. The results demonstrate that explicitly modeling within-curve dependence can substantially improve clustering stability and accuracy, particularly in settings with strong dependence and high noise levels.

Predictive performance at different time horizons in the presence of competing risks: machine learning versus statistical time-to-event models

Josline Otieno

Abstract

Few studies have evaluated competing risk models at multiple time points despite the cumulative incidence of the outcome of interest changing throughout the observation time. This study focused on evaluating the performance of traditional statistical models (the cause-specific Cox and Fine-Gray models), a tree-based model (random survival forests for competing risks), a deep-learning-based model (DeepHit) and pseudo-observation-based models (linear regression, elastic-net linear regression, and random forests models) for competing risk prediction across multiple evaluation time points. The evaluation measures included time-dependent C-index, the integrated calibration index, and the Brier score. The analyses were based on two datasets of different sizes, the Primary Biliary Cirrhosis dataset, for benchmarking, and a large dataset from the Swedish stroke register (Riksstroke). All models indicated similar performance pattern across the evaluation times in both datasets. Model discrimination improved from short term to midterm followed by a decline, while overall prediction accuracy consistently worsened over time. The study concluded that the performance of competing risk models varies substantially across evaluation time points and datasets, demonstrating the need to report performance at multiple time points to guide the appropriate model selection.

Finding the most likely network given observed disease transmissions

Lars Rönnegård and Hector Marina

Abstract

Our research is motivated by data collected on cows inside a dairy barn. We have records on five different contact networks computed using a real-time location system during several weeks coupled with data on mastitis pathogens. We know the strain of each pathogens and have been able trace the transmission of pathogens between individuals. Given the recorded transmissions and the network data, the aim of our study was to develop a likelihood-based method to find the most likely network where the transmission occurred. For simplicity, the calculation of the likelihood assumes that the probability of being newly infected is independent given the network information and the information on who was previously infected. So far we have assessed the model using simulations and the results look promising. We are able to distinguish between networks and find the most likely one. The method should be possible to extend to other transmitted pathogens and other species including humans.

Measuring episodic memory at different time points with different instruments

Ye Tian, Anders Lundqvist and Marie Wiberg

Abstract

To measure episodic memory, several instruments can be implemented repeatedly to capture time-varying changes for certain groups of people, which are treated as longitudinal data. Such health instruments have been increasingly applied with linking or equating methods to make connections between different instruments measuring similar constructs in clinical settings. However, these methods could also be applied when some individual tests are not administered at every time point, i.e., how to scale the new scores with the old scores. Methods to achieve score comparability, a process known as test score equating, often rely on including common test items or assuming that test-taker groups are similar in key characteristics. If such common items are not available, individuals' background information, such as demographic covariates and related test scores, could also serve as anchors, and a propensity score could be introduced to compress information from the covariates. In this study, we include a more conservative scenario, in which even such background information is not available, and how to use historical scores in longitudinal data to obtain the linking transformation. The methods are evaluated using data from the Betula study and through a simulation study. The results show that historical scores could be adopted to replace background covariates with well-balanced and overlapping propensity score functions. Practical implications are discussed, and a workflow is provided for illustration.

The Bachelor's Program in Statistics and Data Analysis at Linköping University: Forty Years of Education and Current Challenges

Annika Tillander

Abstract

The Bachelor's program in Statistics and Data Analysis at Linköping University has been a cornerstone of statistical education in Sweden for over forty years. This poster presents the program's structure, its evolution to meet modern demands, and current challenges such as student retention and adapting to new technologies like large language models. We also provide examples of job titles held by our alumni, illustrating the diverse roles for which the program prepares students. Our experience demonstrates how a long-standing program can balance tradition with innovation to remain relevant in today's data-driven world.

Disaggregation of Swedish households total electricity consumption: Insights for energy efficiency and crisis management

Denise Uwamariya and Vera van Zoest

Abstract

Geographical tensions and increased number of sabotages has led to an increased need for understand- ing where we can reduce electricity consumption in times of an energy crisis. Smart energy meters are being deployed in households to monitor electricity consumption at a high temporal resolution, enabling electricity consumers to manage their consumption, and suppliers to efficiently manage billing and supply. However, typical electricity meters installed in households only provide an aggregate of the household's consumption at the main fuse, not providing information about the consumption of individual appliances. This paper proposes a statistical method to disaggregate household's electricity consumption data from smart meters to understand how electricity is consumed at appliance-level. We demonstrate the method using a Swedish dataset containing smart meter data from 10, 609 households. The results show that throughout the week most electricity is consumed for heating and for kitchen appliances. Comparing our results to previous literature, we see that lighting has become more efficient, while we see an increase in the use of electronic and entertainment appliances compared to previous years. The results of the study will help in targeted communication to citizens in time of an energy crisis, to effectively reduce electricity consumption in an effort to avoid large-scale blackouts. The results also give insight into the type of appliances where we can gain most in energy efficiency development.