Preliminary Program

Statistical Reasoning and Scientific Error

LMU Munich, 1-4 July 2019

1 July, 2019

Time	Event
12:30 – 14:00	Registration and Lunch
14:00 – 14:45	Aris Spanos : The Replication Crises and the Trustworthiness of Empirical Evidence in Economics
14:45 – 15:00	Konstantin Genin and Kevin Kelly: Simplicity, Progress and Replication
15:30 – 16:00	Coffee Break
16:00 – 16:45	Aydin Mohseni: Implications of Soundness-Dependent Effects for Interventions in the Replication Crisis
16:45 – 17:00	Glenn Shafer: Let's Replace p-Values with Betting Outcomes
17:15 – 18:30	Keynote: Deborah Mayo (Virginia Tech): TBA

2 July, 2019

Time	Event
09:00 – 09:45	Michał Sikorski and Mattia Andreoletti: Epistemic and Social Functions of Replicability
09:45 – 10:30	Insa Lawler and Georg Zimmermann: Misalignment Between Research Hypotheses and Statistical Hypotheses – A Threat to Evidence-based Medicine?
10:30 – 11:00	Coffee Break
11:00 – 11:45	Alessandra Cenci and M.Azhar Hussain: Robustness and Equity in Evidence-based Policy
11:45 – 12:30	Aline Claesen, Tom Heyman, Francis Tuerlinckx and Wolf Vanpaemel: The Relation Between Reporting Errors and Data Sharing
12:30 – 14:00	Lunch Break
14:00 – 14:45	Arianne Herrera-Bennett, Chia Wei Ong and Moritz Heene: Exploring Indices of Repeated k-Fold Cross-validation as Predictors of Study Replicability
14:45 – 15:30	Reid Dale: Formally Learning from Error
15:30 – 16:00	Coffee Break
16:00 – 16:45	Amanda Kvarven, Eirik Strømland and Magnus Johannesson: Comparing Meta-Analyses and Pre-Registered Multiple Labs Replication Projects
16:45 – 17:00	Austin Due: Guarantees and Increase: Incentivizing Replications and QRP-Avoidance without Impinging on Inertia and Creativity
17:15 – 18:30	Keynote: Regina Nuzzo (Gallaudet University): TBA

3 July, 2019

Time	Event
09:00 – 09:45	Mark Colyvan: The Role of Toy Statistical Models in Legal Reasoning
09:45 – 10:30	Pavel Janda and Rafal Urbaniak: Probabilistic Models of Legal Corroboration
10:30 – 11:00	Coffee Break
11:00 – 11:45	TBA

11:45 – 12:30	Johannes Keller: Model Selection Arguments for Instrumentalism
12:30 – 14:00	Lunch Break
14:00 – 14:45	Adrian Ziółkowski: On How Incomplete Reporting Affects Replicability. A Case Study from Experimental Philosophy: Failed Replications of Swain et al. (2008)
14:45 – 15:30	Barbara Osimani: Science as a Signaling Game: Statistical Evidence in Strategic Environments
15:30 – 16:00	Coffee Break
16:00 – 16:45	Felipe Romero and Jan Sprenger. Scientific Self-Correction: The Bayesian Way
16:45 – 17:00	William Peden: John Norton, Direct Inference, and Calibrated Bayesianism
17:15 – 18:30	Keynote: Miklós Rédei (LSE): TBA

4 July, 2019

Time	**Event**
09:15 – 10:00	Lorenzo Casini, Mattia Andreoletti and Jan Sprenger: Meta-Analysis and Conflicts of Interest
10:00 – 10:45	Gerit Pfuhl: CRAZED Research? On Epistemic and Instrumental Irrationality in Research
10:45 – 11:15	Coffee Break
11:15 – 12:30	Keynote: Uri Simonssohn (Universitat Ramón Llull): Uri Simonssohn (Universitat Ramón Llull): Rethinking Interactions: Most Published Interactions Have Been Misinterpreted
12:30 – 14:00	Lunch Break
14:00 – 14:45	Mariusz Maziarz: The Use of Inconsistent Causal Inferences from Observational Data for Policymaking
14:45 – 15:30	Lee Jussim : Theoretical and Statistical Misinterpretations of “Implicit Bias”
15:30 – 16:00	Coffee Break
16:00 – 16:45	Adam Kubiak: Socio-cognitive Strategies for Justification of Neyman’s “Inductive Behavior” Conception of The Objective of Science
16:45 – 17:00	David Watson: The Explanation Game: A Formal Framework for Explainable Artificial Intelligence

Abstracts:

Lorenzo Casini (University of Geneva), Mattia Andreoletti and Jan Sprenger (University of Turin): Meta-Analysis and Conflicts of Interest

In medical research, meta-analyses over multiple randomized controlled trials (RCTs) are praised for mitigating the problem of confounding due to the small sample size of individual RCTs. An underestimated limitation of meta-analyses is that many RCTs suffer from conflicts of interest (CoI), raising the question of whether studies with known CoIs should be discounted in this analysis. In this project, we rebut an argument by Fuller (2018, Philosophy of Science) on the relevance of such meta-evidence and investigate whether, and under which conditions, CoIs should affect our statistical conclusions.

Alessandra Cenci (University of Southern Denmark) and M.Azhar Hussain (University of Sharja): Robustness and Equity in evidence-based policy

Replicability of experimental results and frequent errors are major problems in many research fields and for science-related policies. What is argued is that a more extended use of recently developed "robust methods" for economic/health evaluation could be helpful to attain vital scientific and societal goals once the “robust knowledge” is implemented at public policy level. Particularly, their formal properties and certain operational advantages reduce unpredictability and enhance cogency of results. Likewise, they have an intrinsic capacity to support/embody epistemic and non-epistemic values (e.g., evidence, fairness-equity). All this would be crucial in public health analysis informing evidence-based but also equity-oriented policies.

Aline Claesen, Tom Heyman, Francis Tuerlinckx and Wolf Vanpaemel (Katholieke Universiteit Leuven): The Relation Between Reporting Errors and Data Sharing

Sharing one’s research data is still not as common in psychology as it should be, despite its benefits. One possible explanation is that authors fear that reanalysis will not confirm their own conclusions and errors will be exposed. As of now, two empirical studies assessed the relationship between data sharing and inconsistencies in reported test statistics, degrees of freedom and p-values, and their findings differ (Nuijten et al., 2017; Wicherts, Bakker and Molenaar, 2011). This presentation will discuss preliminary findings of a replication of Wicherts, Bakker and Molenaar (2011), focusing on 394 articles instead of the original 49 articles.

Mark Colyvan (University of Sydney): The Role of Toy Statistical Models in Legal Reasoning

A great deal of theorising about the proper place of statistical reasoning in the courtroom revolves around several canonical thought experiments that invoke toy statistical models of the situation in question. I will argue that all of these canonical thought experiments are flawed in various (albeit interesting) ways. In some cases the flaws involve subtle underspecification that leads to ambiguity about the intuitive judgement; in other cases the flaw is that the thought experiment stipulates that we forgo freely-available and relevant evidence. The upshot is that these thought experiments do not succeed to undermine the use of statistical evidence in the courtroom.

Reid Dale (University of California, Berkeley): Formally Learning from Error

Mayo and other error statisticians advocate for severe testing as the appropriate framework guiding the use of statistical methods. In this talk, we analyze the error statistical position from the perspective of Formal Learning Theory. By attempting to faithfully formalize the notion of severe testing, we find that formal logical considerations cast doubt on the virtues of stability and non-comparativity argued for by Mayo. After adopting a revised definition of severe testing we may ask which hypotheses are amenable to error-theoretic analysis. We show that, unlike the Bayesian, hypotheses of high quantifier complexity hypotheses are not severely testable.

Austin Due (Institute for the History and Philosophy of Science and Technology, University of Toronto): Guarantees and Increase: Incentivizing Replications and QRP-Avoidance without Impinging on Inertia and Creativity

Ending the Replication Crisis requires addressing (i) the proliferation of false-positives due to questionable practices and (ii) the lack of replications being performed. Proposals on the market fail to sufficiently address both these points, and future proposals that should address both points must fit within constraints to be considered realistic. These constraints are the perceived ability for scientific creativity and the conservative nature of scientific change. Given (i) and (ii) in light of these constraints, I propose two policies for journals with which a replication is rewarded with a guaranteed future publication and the increase of journal acceptance rates, respectively.

Konstantin Genin and Kevin Kelly (University of Toronto): Simplicity, Progress and Replication

Say that a method for answering a question is progressive if the chance of outputting the true answer increases with sample size. Surprisingly, many standard statistical methods are not even approximately progressive. That amounts to a designed-in tendency toward replication failure. We prove that it is often possible to approximate progressiveness arbitrarily well. Furthermore, every approximately progressive method must obey a version of Ockham's razor. That answers the following question: is there a non-circular justification for Ockham's razor when the truth may well be complex? We demonstrate applications in null-hypothesis statistical testing and in causal discovery from non-experimental data.

Arianne Herrera-Bennett, Chia Wei Ong and Moritz Heene (LMU Munich): Exploring Indices of Repeated k-Fold Cross-validation as Predictors of Study Replicability

The project aims to assess the link between model generalizability and effect replicability: Re-analysis (5-repeated 10-fold cross-validation) of N=21 replication studies (Camerer et al., 2018) will allow us to correlate replication success with model generalizability indices (R-squared, root-mean-squared error (RMSE), mean absolute error (MAE), averaged across folds). Preliminary results align with expectations: R-squared correlated positively ($r = .646$) whereas error indices (i.e. RMSE and MAE) correlated negatively ($rs = -.465$) with replication success. While resampling within the same data set cannot and should not replace independent replications, it may provide a link between out-of-sample generalization and replicability of observed findings.

Pavel Janda and Rafal Urbaniak (University of Gdanks): Probabilistic Models of Legal Corroboration

The aim is to develop a sensible probabilistic model of legal corroboration in response to an attack on the probabilistic approach to legal reasoning due to Cohen. One of Cohen's arguments is that there is no probabilistic measure of evidential support which satisfactorily captures the situation in which independent witnesses testify to the truth of the same proposition (the phenomenon called corroboration). We investigate the properties of several probabilistic measures discussed by Cohen, discuss Cohen's criticism, and develop our own. Finally, we offer a probabilistic measure of corroboration that evades the critical points raised against the ones discussed so far.

Lee Jussim (Rutgers University): Theoretical and Statistical Misinterpretations of “Implicit Bias”

“Implicit bias” has escaped the lab and is now presented in mainstream media, internet memes, political campaigns, and organizational trainings as if it is an established scientific fact. In contrast, I argue that almost nothing about implicit bias is known with anything remotely resembling scientific certainty: There is no consensual scientific definition as to what constitutes implicit bias; 2. A definition of implicit bias offered by founder of the implicit association test (IAT) Greenwald (2017) as describing what it meant to scientists for the prior 20 years is shown to be logically incoherent and empirically unjustified; 3. Exactly what the IAT measures remains unclear, even after 20 years of research; 4. Estimates of the proportion of people who have implicit racial prejudices appear to be wildly overstated. Nonetheless, meta-analyses have shown that IAT scores predict discrimination to at least a modest extent. Next, alternative explanations for gaps are briefly reviewed, highlighting that IAT scores offer only one of many possible such explanations. We then present a series of heuristic models that assume that IAT scores can only explain what is left over, after accounting for other explanations of gaps. This review indicates that it is likely that IAT scores explain only a modest portion of those gaps. Put differently, this review indicates that, even if the IAT fully captures implicit biases, and those implicit biases were completely eliminated, the extent to which racial gaps would be reduced is minimal. I conclude by arguing that, even after 20 years, much more research is needed to understand what the IAT measures and explains with any certainty. If there is sufficient time, I will discuss why scientists sometimes leap to unjustified conclusions, and practices that can limit the likelihood of doing so.

Johannes Keller (LMU Munich/MCMP): Model Selection Arguments for Instrumentalism

I examine scenarios in predictive modeling that favor Elliott Sober's model selection account of instrumentalism. Building on earlier work of Mikkelson (2006), I characterize the circumstances in which simpler, false statistical models predict better than more complex models. By a computer simulation, I identify four influencing factors: the magnitude of parameter values, noise, sample size and correlations among variables. Manifestation of these factors affect, whether instrumentalist arguments against two variants of the no-miracles-argument apply. Finally, I offer a theoretical justification for the results by appealing to results in statistical methodology and discuss their relevance for an epistemology of data science.

Adam Kubiak (Optimum Pareto Foundation): Socio-cognitive Strategies for Justification of Neyman’s “Inductive Behavior” Conception of The Objective of Science

Jerzy Neyman, a co-founder of frequentist paradigm in statistics, dismissed any type of philosophical school which maintained that scientific inference forms a basis for establishing what we should believe. He called the approach the ‘inductive behavior’ philosophy of scientific method. We investigate the issue of whether it is really pointless to use science as a belief regulator and is the principal role of science really to guide actions rather than beliefs, but from cognitive and societal perspective. We provide arguments for positive answers to both of these questions by offering strategies of argumentation other than the meta-mathematical.

Amanda Kvarven, Eirik Strømland (University of Bergen) and Magnus Johannesson (LSE): Comparing Meta-Analyses and Pre-Registered Multiple Labs Replication Projects

Many researchers rely on meta-analysis to summarize research evidence. However, there is a concern that biases in primary studies will carry over into meta-analyses. We compare the results of meta-analyses to large-scale pre-registered replications in psychology. Searching the literature, 17 meta-analyses – spanning more than 1,200 effect sizes and more than 370,000 participants - on the same topics as multiple labs replications are identified. The meta-analytic effect sizes are significantly different from the replication effect sizes for 12 out of the 17 meta-replication pairs. On average, meta-analytic effect sizes are about three times larger than the replication effect sizes.

Insa Lawler (Ruhr University Bochum) and Georg Zimmermann (Paracelsus Medical University & Paris Lodron University of Salzburg): Misalignment Between Research Hypotheses and Statistical Hypotheses – A Threat to Evidence-based Medicine?

Evidence-based medicine uses statistical hypothesis testing. In this paradigm, one tests the negation of a statistical hypothesis (SH) that corresponds to the research hypothesis (RH). Yet in practice, misalignments between RH and SH frequently occur, e.g., directional RHs are paired with non-directional SHs. In our paper, we (i) specify different forms of misalignments, (ii) provide reasons for the occurrence of misalignments, (iii) argue that the available counterbalances do not cover all cases and lead to methodological inadequacy, loss of statistical power, or a (potential) lack of information that could be crucial for clinical decisions, (iv) suggest some remedies.

Mariusz Maziarz (Wroclaw University of Economics): The Use of Inconsistent Causal Inferences from Observational Data for Policymaking

The results of observational studies (the research typical to epidemiology and econometrics) lack stability. The presence of recalcitrant results undermines informing theoretical discourse and putting forward policy guidance. We show that the traditional approaches to inference from empirical literature such as meta-analysis and QATs are not useful. Instead, we offer an alternative approach based on the notion of ‘extrapolatory distance’: inconsistent results can be interpreted as proxies for/approximations of different policy settings and conclusions should be based on the ground of most relevant study for a given context instead of a meta-analysis averaging inconsistent estimates of the treatment effect.

Aydin Mohseni (University of California, Irvine): Implications of Soundness-Dependent Effects for Interventions in the Replication Crisis

Scientific studies vary in their methodological soundness. Interventions in evidentiary standards and research practices can differentially affect studies as a function of their soundness. The conjunction of these facts has unrecognized implications for proposed interventions in the replication crisis. I argue that would should expect these facts to obtain, and demonstrate that, when accounting for differential effects of interventions as a function of soundness, several of the proposed interventions---lowering the significance threshold, promoting preregistration, and sample splitting---will produce less improvement than

estimates would suggest and, in some cases, actually increase false discovery rates, sign error rates, and magnitude exaggeration ratios.

Barbara Osimani (LMU Munich/MCMP): Science as a Signaling Game: Statistical Evidence in Strategic Environments

As a response to the “reproducibility crisis” and to a general crisis of trust towards the scientific enterprise (Edwards and Roy 2017, Vazire 2017), various initiatives are being promoted in order to foster transparency (see e.g. Open Science Movement, AllTrials Campaign, Sense about Science). We advance that game theory should be used to explain different kinds of biases and identify solutions to them. We will focus on two specific settings: interactions between the pharmaceutical industry and authorities that regulate drug approval (as well as other components of the medical ecosystem) and scientific publication systems.

William Peden (Universita Politecnica delle Marche): John Norton, Direct Inference, and Calibrated Bayesianism

John Norton has been one of the most prominent critics of Bayesian philosophy of induction in recent years, via his "Material Theory of Induction" (MTI). While he has not discussed statistical inference in depth, his theory has intriguing implications for the foundations of statistics. Norton believes that local factual beliefs lie at the heart of induction, and these belief states can only sometimes be probabilistically represented. Neither entropy maximization nor betting odds play a fundamental role in science. I raise two problems for the MTI; I also argue it should be seen as complementary rather than a competitor to Bayesianism.

Gerit Pfuhl (UiT The Arctic University of Norway): CRAZED Research? On Epistemic and Instrumental Irrationality in Research

I propose to group scientific errors into epistemic and instrumental. Epistemic irrationality can be mitigate by discourse, relying on our tendency to be critical towards the argument from others. Gradually, beliefs that are not supported with scientific evidence and that do not hold up to scrutiny via replications, will disappear and be replaced by newer beliefs and theories. The capacity for logical and deliberate reasoning is necessary but not sufficient for an agent to act accordingly. It can be, given certain incentive structures, rational to act against ones epistemic belief. Accordingly, instrumental irrationality requires tailored strategies and changes in incentives.

Felipe Romero (RU Groningen) and Jan Sprenger (University of Turin): Scientific Self-Correction: The Bayesian Way

There are different approaches for addressing the replication crisis in science. Social reformists hypothesize that the social structure of science such as the credit reward scheme must be changed. Statistical reformists argue more specifically that science would be more reliable and self-corrective if null hypothesis significance tests (NHST) were replaced by a different inference framework, such as Bayesian statistics. On the basis of a simulation study for meta-analytic aggregation of effect sizes, we articulate a middle ground between the

different reform proposals: statistical reform alone won't suffice, but moving to Bayesian statistics eliminates important sources of overestimating effect sizes.

Glenn Shafer (Rutgers University): Let's Replace p-Values with Betting Outcomes

We can think of a Neyman-Pearson 5% test as a bet that multiplies your money by 0 or 20. Suppose that instead of measuring the strength of evidence with a single N-P test or a p-value, you make a bet that can multiply the money you risk by many different factors. The factor by which it does multiply your money measures the strength of the evidence. This leads to replacements for the concepts of power and confidence and to methods for meta-analysis and multiple testing. It supports an understanding of objective probability that avoids the notion of unseen alternative worlds.

Michał Sikorski and Mattia Andreoletti (Universitá degli Studi di Torino): Epistemic and Social Functions of Replicability

Is replicability a crucial feature of science? Many philosophers of science have discussed the limitations rather than the advantages of replicability (see e.g. Leonelli 2018; John Norton 2015). Whereas, scientists see replicability as one of the defining features of their disciplines and consider the high rate of replicability failures as an “apocalypse” for science (Bishop, 2019). We will try to make sense of this tension. We start by specifying what replicability means. We will defend the epistemic and social value of replicability. Finally, we will suggest a strategy to select the appropriate level of replications and present case studies.

Uri Simonsohn (Universitat Ramón Llull): Rethinking Interactions: Most Published Interactions Have Been Misinterpreted

Hypotheses involving interactions are common in social science. Do returns to education differ by gender? Are unexpected losses more impactful than expected ones? Does construal moderate power posing? Etc. Linear regressions, as in $y=\boldsymbol{a}x+\boldsymbol{b}z+\boldsymbol{c}xz$ are the most common (only?) way in which such interactions are tested, and yet such approach is extremely likely to give the wrong answer. The false-positive rate can easily reach over 50%, the sign of the interaction can easily be wrong, the average interaction effect is oddly defined and its estimate often biased, and simple-slopes/spotlight/floodlight analysis is approximately hopeless. I identify four questions we often ask from interaction effects, and explore trustworthy alternatives to the current approach to answering them.

Aris Spanos (Virginia Tech): The Replication Crises and the Trustworthiness of Empirical Evidence in Economics

It is argued that the abuse of significance testing is only a symptom of a much broader problem relating to the uninformed application of statistical methods without real understanding of their assumptions, proper implementation and cogent interpretation of their inferential results. The paper makes a case that the trustworthiness of empirical evidence should be assessed at the individual study level, and not at a discipline-wide level. It is argued that the three most important sources of untrustworthy evidence are: (i) statistical misspecification: invalid probabilistic assumptions imposed on one's data, (ii) poor implementation of inferential methods, and (iii) unwarranted evidential interpretations.

Vipul Vivek (Jawaharlal Nehru University): Values in Science and Aleatory Uncertainty

Douglas 2000 argues objectivity in science is impractical as non-epistemic values are necessary wherever non-epistemic consequences of error exist. I propose to move the focus of this debate to responsibility at least in relation to aleatory uncertainty (intrinsic variability) as opposed to epistemic uncertainty (lack of knowledge). Uncovering what Winsberg 2012 calls 'uncorrectibly involved social and ethical values' in science cannot help reduce aleatory uncertainty. It would be better to simply make it explicit through uncertainty quantification. However current statistical frameworks collapse this distinction, reducing the variety in risk attitudes policymakers could have had if the distinction were known.

David Watson (University of Oxford): The Explanation Game: A Formal Framework for Explainable Artificial Intelligence

I propose a formal framework for explainable artificial intelligence (XAI). Combining elements from epistemological pragmatism, statistical learning, and game theory, I design an idealised explanation game in which players collaborate to find the best explanation for a given algorithmic prediction. Through an iterative procedure of questions and answers, the players establish a three-dimensional Pareto frontier that defines the optimal trade-offs between explanatory accuracy, simplicity, and relevance. I characterise the conditions under which such a game is almost surely guaranteed to converge on an optimal explanation surface in polynomial time, and illustrate my proposal with a number of real-world examples.

Adrian Ziółkowski: On How Incomplete Reporting Affects Replicability. A Case Study from Experimental Philosophy: Failed Replications of Swain et al. (2008)

The paper focuses on three replication attempts of a study originally conducted by Swain et al. (2008), whose results received much recognition in the philosophical literature. We present data that do not corroborate the original findings and provide an in-depth discussion of factors we observed in the process that negatively affect the replicability of the original experiment. We will use this case study to illustrate the importance of precise reporting (concerning both data and methodological or procedural details) for the replicability of experimental studies. We will also put forward few hypotheses why the original study tends not to replicate.