Emerging Challenges for Statistics and Data Sciences: Complex Data with Missingness, Measurement Errors, and High Dimensionality

Shu Yang (North Carolina State University),
David Haziza (University of Ottawa)
Peng Ding (University of California, Berkeley)
Chenyin Gao (North Carolina State University)
Grace Yi (Western University)

May 22 - May 27, 2022

The workshop "Emerging Challenges for Statistics and Data Sciences: Complex Data with Missingness, Measurement Errors, and High Dimensionality" was held at The University of British Columbia - Okanagan Campus from May 22 to May 27 in a hybrid manner due to COVID 19 pandemic. The workshop's main objective is to connect people with similar research interests by fostering in-depth discussions. Moreover, It allows the leading experts to present their cutting-edge research products by showcasing the state-of-the-art methodologies related to missingness, measurement errors, and high dimensionality, eventually driving future academic collaborations.

1 Overview of the Field

Advanced by modern technology, statistical science has gone through a paradigm shift from traditional statistics to using big data and modern machine learning. For example, recently, the integration of gold-standard probability samples and emerging big non-probability samples has been a popular research topic in survey statistics. As promising as it is, big data also presents inevitable challenges such as heterogeneous data sources, selection bias, missingness, measurement errors, and high-dimensionality, which requires principled approaches. Essential to statistics are quantitative methods for translating complex data into meaningful information and knowledge to guide policy and decision-making to reach new scientific discoveries.

The past years have seen tremendous progress in research advances in new theories, methods, and algorithms from different branches of statistics for surmounting fundamental challenges arising from newly emergent big data. It is therefore crucial and timely to bring statisticians, data and computer scientists, and practitioners together to share recent advances, identify pressing problems, and spark productive collaborations, ultimately leading to fundamentally new methodological advances, accelerating the progress of using big data to answer critical scientific questions.

In this workshop, we invited several speakers to talk about their research in data integration in survey sampling and causal inference, non-ignorable and high-dimensional missingness, and novel statistical methodologies that cope with data in various formats. Each talk highlights important real-life applications that could arise in many fields, including official statistics, treatment evaluation, precision medicine, and public health, facilitating the critical and informative interdisciplinary exchange.

2 Recent Developments and Open Problems

The invited speakers covered a wide range of recently developed methodologies and raised several open challenges for future research directions. It is enlightening to review their insightful presentations and fruitful interactions, in particular:

1. Integrative analysis of probability sample and non-probability sample.
2. Generalizability and transportability from clinical trial to target populations.
3. Missing imputation coupled with nonignorability or high-dimensional data.
4. Causal graphic models.
5. Personalized treatment regimes.
6. Novel methods dealing with unconventional structured data.

3 Presentation Highlights

This workshop included 28 talks, with five talks delivered in-person and the rest over Zoom. Each talk is scheduled for 45 minutes, including a 5-10 minutes discussion. In addition, coffee breaks were considered every two talks for informal discussions among the in-person participants. As a result, our BIRS workshop achieved high levels of demographic diversity among the speakers; see Figure 1 for summaries.

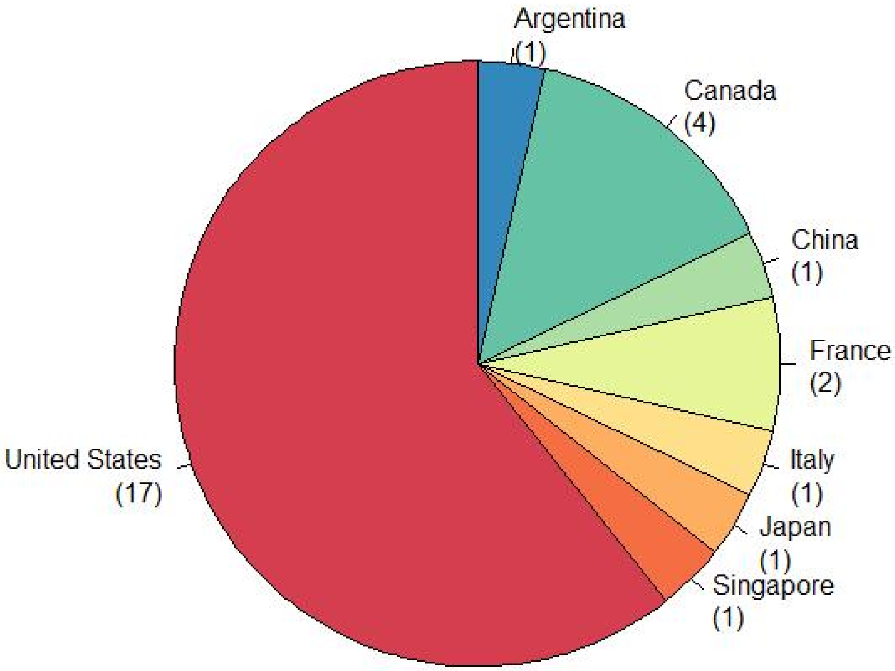

Figure 1: Demographic distribution of workshop speakers

Synopses of the invited talks are presented below.

Day 1. The workshop started with the presentation by *Andrea Rotnitzky* (Di Tella University) on Monday (May 23), titled “Towards deriving graphical rules for efficient estimation in causal graphical models”. Her talk is described as:

Causal graphs are responsible in great part for the explosive growth and adoption of causal inference in modern epidemiology and medical research. This is so because graphical models facilitate encoding and communicating causal assumptions and reasoning in an intuitive fashion, requiring minimal, if any at all, mathematical dexterity. Applying simple graphical rules, it is possible to easily communicate biases due to confounding and selection, which data are needed to correct for these biases, and to derive which statistical target parameter quantifies a causal effect of interest. Yet, causal graphical models encode a well-defined statistical semiparametric model, and little, if any, work has been done to investigate and derive simple graphical rules to encode efficiency in estimation. In this talk, I will present work toward bridging this gap. Given a causal graphical model, I will derive a set of graphical rules for determining the optimal adjustment set in a point exposure problem. This is the subset of variables in the graph that both suffices to control for confounding under the model and yields a non-parametric estimator of the population average treatment effect (ATE) of a dynamic, i.e. personalized, or static point exposure on an outcome with the smallest asymptotic variance under any law in the model. I will then discuss the conditions for the existence of a universally optimal adjustment set when the search is restricted to a realistic scenario in which only a subset of the graph variables is observable. For the problem of estimating the effect of a time-dependent treatment, I will discuss an impossibility result. Finally, I will summarize recent results on graphical rules for constructing a reduced graph whose nodes represent only those variables that are informative for ATE and such that efficient estimation of ATE under the reduced graph and under the original graph agree.

In the following, *Fan Li* (Duke University) presented her recent work with the name ”Are deep learning models superior for missing data imputation in surveys? Evidence from an empirical comparison” and the detailed abstract are as follows:

Multiple imputation (MI) is a popular approach for dealing with missing data arising from non-response in sample surveys. Multiple imputation by chained equations (MICE) is one of the most widely used MI algorithms for multivariate data, but it lacks theoretical foundations and is computationally intensive. Recently, missing data imputation methods based on deep learning models have been developed with encouraging results in small studies. However, there has been limited research on evaluating their performance in realistic settings compared to MICE, particularly in large-scale surveys. We conduct extensive simulation studies based on American Community Survey data to compare the repeated sampling properties of four machine learning-based MI methods: MICE with classification trees, MICE with random forests, generative adversarial imputation networks, and multiple imputation using denoising autoencoders. We find the deep learning-based imputation methods are superior to MICE in terms of computational time. However, with the default choice of hyperparameters in the common software packages, MICE with classification trees consistently outperforms, often by a large margin, the deep learning imputation methods in terms of bias, mean squared error, and coverage under a range of realistic settings. This is a joint work with Zhenhua Wang, Jason Poulos and Olanrewaju Akande.

After a short coffee break, the workshop continued with *Jae-Kwang Kim* (Iowa State University) presenting “An information projection approach to propensity score estimation for handling selection bias in voluntary samples”. His talk abstract is listed as follows:

Propensity score weighting is widely used to improve the representativeness and correct the selection bias in the sample. The propensity score is often developed using a model for the sampling probability, which can be subject to model misspecification. We consider an alternative approach of estimating the inverse of the propensity scores using the density ratio function. The smoothed density ratio function is obtained by the solution to the information projection onto the space satisfying the moment conditions on the balancing scores. By including the covariates for the outcome regression models only in the density ratio model, we can achieve efficient propensity score estimation. Extension to nonignorable selection mechanism will be discussed.

Next, *Jörn Diedrichsen* (Western University) presented his work with the name “Estimating human brain organization by fusion across functional imaging datasets”. In his talk abstract, it is described as:

Functional magnetic resonance imaging (fMRI) allows the simultaneous observation of activity in hundreds of thousand brain locations (voxels), while human participants can be engaged in a large variety of mental tasks. The resultant datasets have been used to produce functional atlases of the human brain, usually by subdividing the brain into a finite set of discrete regions, each associated with a specific function. In recent years, the number of high-quality data sets and the number of associated brain parcellation maps have rapidly increased. However, each individual dataset typically focuses on a specific functional domain, often leading to poor characterization of other parts of the brain. In my talk, I will present a framework that allows the fusion of disparate functional data sets to produce a more complete characterization of the human brain. The backbone of this framework is a high-dimensional probabilistic model that describes the spatial arrangement of functional brain regions. Simultaneously, the framework learns separate emission models for each data set, each one linking the functional brain regions to the predicted response for the specific set of tasks. The framework integrates information across diverse data sets while taking into account the relative strengths and weaknesses of each efficiently deals with missing data within individual subjects and may help to provide better estimates of individual brain organization based on limited data.

In the afternoon, *Camelia Goga* (Université de Franche-Comté) presented her recent work titled "Random-forest model-assisted estimation in finite population sampling" with abstract as:

In surveys, the interest lies in estimating finite population parameters such as population totals and means. In most surveys, some auxiliary information is available at the estimation stage. This information may be incorporated in the estimation procedures to increase their precision. In this article, we use random forests to estimate the functional relationship between the survey variable and the auxiliary variables. In recent years, random forests have become attractive as National Statistical Offices have now access to a variety of data sources, potentially exhibiting a large number of observations on a large number of variables. We establish the theoretical properties of model-assisted procedures based on RFs and derive corresponding variance estimators. A model-calibration procedure for handling multiple survey variables is also discussed. The results of a simulation study suggest that the proposed point and estimation procedures perform well in terms of bias, efficiency, and coverage of normal-based confidence intervals, in a wide variety of settings.

After that, *Paul Zivich* (the University of North Carolina at Chapel Hill) gave a talk about "Fusion to address systematic errors across data sources" and his abstract is presented here:

Fusion study designs seek to combine data from different sources to answer questions that could not be as adequately answered by any single subset. As a didactic example, we consider the identification and estimation of the mean of a covariate for a well-defined population using a conjunction of data sources to address missingness and measurement error. As an elucidatory example, we demonstrate a bridged treatment fusion using historical data from the AIDS Clinical Trial Group. Specifically, we fuse two trials to estimate the risk of a composite outcome (death, AIDS, or greater than a 50% CD4 cell count decline) given triple antiretroviral therapy versus mono antiretroviral therapy through a shared trial arm of dual therapy. Fusion designs, like those illustrated here, allow for principled integration of information across different data sources, help to clarify identification assumptions, and address questions that no single data source could address as well.

Shortly after the coffee break, *Daniel Scharfstein* (University of Utah) presented his work on "Trials with Irregular and Informative Assessment Times: A Sensitivity Analysis Approach" and his work can be summarized in abstract as:

Many trials are designed to collect outcomes at pre-specified times after randomization. However, in practice, there is often substantial variability in the times at which participants are actually assessed, which poses a challenge to learning about the treatment effects at the targeted assessment times. Any method for analyzing a trial with such irregular assessment times relies on untestable assumptions. Therefore, conducting a sensitivity analysis is an important step that can strengthen conclusions from such trials. However, no sensitivity analysis methodology has been developed. We develop a methodology that accounts for possibly informative assessment times, where assessment at time t may be related to the outcome at that time, even after accounting for observed past history. We implement our methodology using a new augmented inverse-intensity-weighted estimator, and we apply it to a trial of low-income participants with uncontrolled asthma. We also evaluate the performance of our estimation procedure in a realistic simulation study. This work is joint with Bonnie Smith, Shu Yang and Andrea Apter.

We regret to miss the talk by *Roderick Little* (University of Michigan) due to our last-minute schedule change. He was scheduled to talk about two key ideas in survey nonresponse, namely response propensity and missing at random. His talk abstract is illustrated as:

I present recent work concerning two key ideas in survey nonresponse, namely response propensity and missing at random. I propose a specific definition of the response propensity that clarifies the conditioning, and weakened sufficient conditions for missing at random for asymptotic frequentist maximum likelihood inference. Finally, I show how an explicit modeling approach allows certain missing not at random mechanisms to be identified when there is post-stratification information.

Day 2. The morning session started with *Joan Hu* (Simon Fraser University) presenting "Statistical Issues on Large Administrative Health Data" and her abstract is provided as:

Administrative health data are primarily generated through routine administrative health programs, and typically record information on a broad range of variables over time. Canada's publicly funded health care system has resulted in large provincial medical insurance and disease registry databases. These databases provide rich information on health care, and are usually readily available upon approval by the authority. There has recently been increased interest in using administrative health data to achieve various scientific goals. This presentation showcases challenges arising from analysis of administrative health data and strategies for addressing them. We will focus on unconventional structures of such data and spatio-temporal correlations underlying the records. Particular discussions will be on inference procedures dynamically adaptive to upcoming data and design considerations for future research projects with administrative data.

After that, *Raymond Wong* (Texas A&M University) presented his recent work on "Matrix Completion with Model-free Weighting". He described his talk as:

In this work, we propose a novel method for matrix completion under general non-uniform missing structures. By controlling an upper bound of a novel balancing error, we construct weights that can actively adjust for the non-uniformity in the empirical risk without explicitly modeling the observation probabilities, and can be computed efficiently via convex optimization. The recovered matrix based on the proposed weighted empirical risk enjoys appealing theoretical guarantees. In particular, the proposed method achieves stronger guarantee than existing work in terms of the scaling with respect to the observation probabilities, under asymptotically heterogeneous missing settings (where entry-wise observation probabilities can be of different orders). These settings can be regarded as a better theoretical model of missing patterns with highly varying probabilities. We also provide a new minimax lower bound under a class of heterogeneous settings. Numerical experiments are also provided to demonstrate the effectiveness of the proposed method.

After a brief coffee break, *Xinran Li* (the University of Illinois at Urbana-Champaign) presented his work titled "Randomization Inference beyond the Sharp Null: Bounded Null Hypotheses and Quantiles of Individual Treatment Effects". The corresponding abstract is here:

Randomization (a.k.a. permutation) inference is typically interpreted as testing Fisher's "sharp" null hypothesis that all effects are exactly zero. This hypothesis is often criticized as uninteresting and implausible. We show, however, that many randomization tests are also valid for a "bounded" null hypothesis under which effects are all negative (or positive) for all units but otherwise heterogeneous. The bounded null is closely related to important concepts such as monotonicity and Pareto efficiency. Inverting tests of this hypothesis yields confidence intervals for the maximum (or minimum) individual treatment effect. We then extend randomization tests to infer other quantiles of individual effects, which can be used to infer the proportion of units with effects larger (or smaller) than any threshold. The proposed confidence intervals for all quantiles of individual effects are simultaneously valid, in the sense that no correction due to multiple analyses is needed. In sum, we provide a broader justification for Fisher randomization tests, and develop exact nonparametric inference for quantiles of heterogeneous individual effects. We illustrate our methods with simulations and applications, where we find that Stephenson rank statistics often provide the most informative results.

Erica Moodie (McGill University) talked about "Penalized doubly robust regression-based estimation of adaptive treatment strategies". She described her talk in abstract as:

Adaptive treatment strategies (ATSs) are often estimated from data sources with many covariates measured, only a subset of which are useful for tailoring treatment or control of confounding. In such cases, including all the covariates in the analytic model could possibly yield an inappropriate or needlessly complicated treatment decision. Hence, it is crucial to apply variable selection techniques to ATSs. Variable selection with the objective of optimizing treatment decisions has been the subject of only very little literature. In this talk, I will present a regression-based estimation method that can naturally incorporate variable selection through a penalization approach that incorporates sparsity while ensuring strong heredity, and show how we can additionally incorporate confounder selection into the approach. We illustrate the methods by analyzing a pilot sequential multiple assignment randomized trial of a web-based, stress management intervention using a stepped-care method for cardiovascular disease patients to determine useful tailoring variables while adjusting for chance imbalances in important covariates due to the smaller sample size in the pilot.

Our afternoon session begins with *Peisong Han* (University of Michigan) presenting "Integrating summary information from many external studies with heterogeneous populations". He described his talk as follows:

For an internal study of interest, the information provided by relevant external studies can be useful to improve the efficiency of parameter estimation in model building, and the external information is oftentimes in summary form. When information is available from possibly many external studies, extra care is needed due to inevitable study population heterogeneity. The information from studies with populations different from the internal study may harm model fitting by introducing estimation bias. We allow the number of external studies that can be considered for possible information integration to increase with the internal sample size, and develop an effective method that integrates only the helpful information for efficiency improvement without introducing bias. Using this method, we study the change of mood symptoms from the pre-COVID pandemic era to the pandemic era for individuals with bipolar disorder, by integrating summary information from relevant existing large-scale studies to improve efficiency.

Fan Yang (University of Colorado Anschutz Medical Campus) gave a talk on "Identifiability of direct and indirect effects in mediation studies with nonignorable missingness in mediator and outcome". The abstract is given as follows:

Mediation analysis is a useful and widely adopted approach for investigating causal pathways through which an effect arises. However, many mediation studies are challenged by missingness in the mediator and/or the outcome. In general, when the missingness is nonignorable, the direct and indirect effects are not identifiable. In this work, we explore and prove the identifiability of the direct and indirect effects under some interpretable nonignorable missingness mechanisms. We evaluate the performance of statistical inference under those mechanisms through simulation studies and use the National Job Corps Study as an illustration.

Shortly after the coffee break, *Kosuke Morikawa* (Osaka University) presented his work on "Semiparametric adaptive estimation under informative sampling". The abstract is given below:

In survey sampling, it is often difficult to equitably sample data from a population so that the samples are biased. However, information on the inclusion probabilities of samples is available to remove the selection bias. The Horvitz-Thompson estimator is one of the most well-known debiased estimators. Although the Horvitz-Thompson estimator is consistent and asymptotically normal, it is not efficient. In this talk, we derive the semiparametric efficiency bound for various target parameters by treating the inclusion probability itself as a random variable and propose a semiparametric optimal estimator with some working models on the inclusion probability. The proposed estimator is consistent, asymptotically normal, and efficient among the regular and asymptotically linear estimators. We apply the proposed method to the 1999 Canadian Workplace and Employee Survey data.

Jay Breidt (NORC at the University of Chicago) ended the day with a talk on "Dual-frame estimation approaches for combining probability and nonprobability samples" and the corresponding abstract is:

In some complex surveys with two stages of sample selection, primary sampling units (PSUs) are screened for secondary sampling units of interest, which are then measured or subsampled. PSUs without secondary units of interest are costly and time-consuming. Motivated by the low yield of the Large Pelagics Intercept Survey, a two-stage screening sample for a rare type of recreational fishing activity, we have considered surveys that allow expert judgment in the selection of some PSUs. This non-probability judgment sample is then combined with a probability sample to generate likelihood-based estimates of inclusion probabilities and estimators of population totals that are related to dual-frame estimators. Properties of the dual-frame estimation technique are described, and the methods are applied to other problems of combining probability and nonprobability sample data, including respondent-driven sampling. In simulation experiments, the estimators show considerable robustness to misspecification of the nonprobability sampling mechanism.

Day 3. Today started with *Jeff Buzas* (University of Vermont) talking about "Relations between margin-based binary classifiers and logistic regression". The talk abstract is illustrated as:

This talk explores new connections between logistic regression and margin-based binary classification methods. The connections provide novel perspectives and insight on classification methods that use exponential loss, logistic loss, and other commonly used loss functions. The connections also suggest new approaches to adjusting for covariate measurement error in logistic regression with the lasso and/or ridge constraints. Additionally, a general class of loss functions is defined with population minimizer interpretable on the logit scale. The class includes exponential, logistic, logistic regression, Savage, and α tunable loss functions, thereby providing additional insight as to their commonalities and differences. An interesting new loss function emerges from the general class. Properties of this new loss function are explored.

Julie Josse (Inria) gave a talk on missing data analysis, titled "Supervised learning with missing values". The abstract is given in below:

An abundant literature addresses missing data in an inferential framework: estimating parameters and their variance from incomplete tables. Here, I will review recent works on supervised-learning settings: predicting a target when missing values appear in both training and testing data. First, we study the seemingly-simple case where the target to predict is a linear function of the fully-observed data and show that multilayer perceptrons with ReLU activation functions can be consistent but highly complex. Based on a Neumann series approximation of the optimal predictor, we propose a new principle architecture, called NeuMiss networks. Their originality and strength come from the use of a new type of non-linearity: multiplication by the mask. We provide an upper bound of the Bayes risk of NeuMiss networks, and we show that they have good predictive accuracy. Then, we go beyond the linear regression setting and show how imputation can be coupled with powerful learners such as random forest to achieve consistency.

After a short break, *Jiwei Zhao* (University of Wisconsin-Madison) gave a talk on "Statistical Exploitation of Unlabeled Data under High Dimensionality", which he described as:

We consider the benefits of unlabeled data in the semi-supervised learning setting under high dimensionality, for parameter estimation and statistical inference. In particular, we address the following two important questions. First, can we use the labeled data as well as the unlabeled data to construct a semi-supervised estimator such that its convergence rate is faster than the supervised estimator? Second, can we construct confidence intervals or hypothesis tests that are guaranteed to be more efficient or powerful than the supervised estimator? We show that the semi-supervised estimator with a faster convergence rate exists under some conditions, and the implementation of this optimal estimator needs a reasonably good estimation of the conditional mean function. For statistical inference, we mainly propose a safe approach that is guaranteed to be no worse than the supervised estimator in terms of statistical efficiency. Not surprisingly, if the conditional mean function is well estimated, our safe approach becomes semi-parametrically efficient. After the theory development, I will also present some simulation results as well as a real data analysis. This is based on a joint work with Siyi Deng (Cornell), Yang Ning (Cornell) and Heping Zhang (Yale).

Sixia Chen (The University of Oklahoma Health Sciences Center) then presented a related talk with the title "Multiple model-assisted approach to missing data in survey sampling: more than multiply robustness". He described his talk as:

Missing data analysis requires assumptions about an outcome model or a response probability model to adjust for potential bias due to nonresponse. Doubly robust (DR) estimators are consistent if at least one of the models is correctly specified. Multiply robust (MR) estimators extend DR estimators by allowing for multiple models for both the outcome and/or response probability models, and are consistent if any of the multiple models is correctly specified. We propose a new class of multiple model-assisted estimators, which is more robust than the existing DR and MR estimators, by relaxing parametric model assumptions for the outcome variable and response probability, where any semiparametric, nonparametric or machine learning models can be used as well. The proposed estimator achieves design unbiasedness by using a subsampling Rao-Blackwell method, given cell-homogenous response, regardless of any working models for the outcome. An unbiased variance estimation formula is proposed, which does not use any replicate jackknife or bootstrap methods. The simulation study illustrates the robustness of our proposed methods compared with other existing methods.

Day 3 afternoon was free for informal group discussion and excursions to Kelowna Downtown.

Day 4. On Thursday (May 26), Professor *Alessandra Mattei* (University of Florence) started with a talk titled "Assessing causal effects in the presence of treatment switching through principal stratification." She described her talk as follows:

Clinical trials often allow patients in the control arm to switch to the treatment arm if their physical conditions are worse than certain tolerance levels. For instance, treatment switching arises in the Concorde clinical trial, which aims to assess causal effects on the time to disease progression or death of immediate versus deferred treatment with zidovudine among patients with asymptomatic HIV infection. The Intention-To-Treat analysis does not measure the effect of the actual receipt of the treatment and ignores the information of treatment switching. Other existing methods reconstruct the outcome a patient would have had s/he not switched under strong assumptions. We re-define the problem of treatment switching using principal stratification, and focus on causal effects for patients belonging to subpopulations defined by the switching behavior under control. We use a Bayesian approach to inference taking into account that (i) switching happens in continuous time; (ii) switching time is not defined for patients who never switch in a particular experiment; and (iii) survival time and switching time are subject to censoring. We apply this framework to analyze synthetic data based on the Concorde study. Our data analysis reveals that immediate treatment with zidovudine increases survival time for never switchers, and that treatment effects are highly heterogeneous across different types of patients defined by the switching behavior.

Then Professor *Zhichao Jiang* (University of Massachusetts) gave a related talk on causal inference with the title "Experimental Evaluation of Algorithm-Assisted Human Decision Making." He described his talk as follows:

Despite an increasing reliance on fully-automated algorithmic decision-making in our day-to-day lives, human beings still make highly consequential decisions. As frequently seen in business, healthcare, and public policy, recommendations produced by algorithms are provided to human decision-makers to guide their decisions. While there exists a fast-growing literature evaluating the bias and fairness of such algorithmic recommendations, an overlooked question is whether they help humans make better decisions. Using the concept of principal stratification, we develop a statistical methodology for experimentally evaluating the causal impacts of algorithmic recommendations on human decisions. We propose the evaluation quantities of interest, identification assumptions, and estimation strategies. We also develop sensitivity analyses to assess the robustness of empirical findings to the potential violation of a key identification assumption. We apply the proposed methodology to preliminary data from the first-ever randomized controlled trial that evaluates the pretrial Public Safety Assessment (PSA) in the criminal justice system.

After a brief coffee break, Professor *Edward Kennedy* (Carnegie Mellon University) gave a talk on "Nonparametric Estimation of Heterogeneous Effects", with the abstract below:

Heterogeneous effect estimation plays a crucial role in causal inference, with applications across medicine and social science. Many methods for estimating conditional average treatment effects (CATEs) have been proposed in recent years, but there are important theoretical gaps in understanding if and when such methods are optimal. This is especially true when the CATE has nontrivial structure (e.g., smoothness or sparsity). This talk surveys work across two recent papers in this context. First, we study a two-stage doubly robust CATE estimator and give a generic model-free error bound, which, despite its generality, yields sharper results than those in the current literature. The second contribution is aimed at understanding the fundamental statistical limits of CATE estimation. To that end, we resolve this long-standing problem by deriving a minimax lower bound, with matching upper bound based on higher-order influence functions.

Then Professor *Rebecca Andridge* (Ohio State University) gave a talk on selection bias, with the title "Measures of Selection Bias for Proportions Estimated from Non-Probability Samples, With Application to Polling Data". She described her talk as follows:

The proportion of individuals in a finite target population that has some characteristic of interest is arguably the most commonly estimated descriptive parameter in survey research. Unfortunately, the modern survey research environment has made it quite difficult to design and maintain probability samples: the costs of survey data collection are rising, and high rates of nonresponse threaten the basic statistical assumptions about probability sampling that enable design-based inferential approaches. As a result, researchers are more often turning to non-probability samples to make descriptive statements about populations. Non-probability samples do not afford researchers the protection against selection bias that comes from the ignorable sample selection mechanism introduced by probability sampling, and descriptive estimates based on non-probability samples may be severely biased as a result. In this seminar, I describe a simple model-based index of the potential selection bias in estimates of population proportions due to non-ignorable selection mechanisms. The index depends on an inestimable parameter that captures the amount of deviation from selection at random; this parameter ranges from 0 to 1 and naturally lends itself to a sensitivity analysis. I illustrate its use via simulation and via application to pre-election polling data from the U.S. 2020 Presidential Election.

After lunch, Professor *Changbao Wu* (University of Waterloo) gave a talk on "Dealing with Under-Coverage Problems for Non-probability Survey Samples". His abstract is below:

We discuss two typical scenarios of under-coverage with non-probability survey samples. We show that existing estimation procedures can be used to handle the scenario of stochastic under-coverage. The other scenario, termed as deterministic under-coverage, presents real challenges for valid statistical inferences to the target population. Calibration methods and split population techniques are shown to be useful to reduce biases due to under-coverage problems when a reference probability survey sample with auxiliary information is available.

Then Professor *Karthika Mohan* (Oregon State University) gave a talk on "Graphical Models and Causality for Handling Corrupted Data". Her view was unique among the speakers of the day. She described her talk as follows:

The quality of data given as input to an algorithm determines the quality of its output. Unfortunately, in the real-world, quality of data is often compromised by problems such as missing values, measurement error, selection bias and interference. This talk will focus on two of these problems: missing data and interference. In particular, this talk will outline how causal graphs can be used to model these problems and derive conditions under which consistent estimates of quantities of interest such as mean of a variable and causal effect can be computed.

After the coffee break, Professor *Anqi Zhao* (National University of Singapore) gave a talk on rerandomization, with the title "No star is good news: a unified look at rerandomization based on p-values from covariate balance tests." Her abstract is below:

Randomized experiments balance all covariates on average and provide the gold standard for estimating treatment effects. Chance imbalances nevertheless exist more or less in realized treatment allocations, subjecting subsequent inference to possibly large variability and conditional bias. Modern social and biomedical scientific publications require the reporting of covariate balance tables with not only covariate means by treatment group but also the associated p-values from significance tests of their differences. The practical need to avoid small p-values renders balance check and rerandomization by hypothesis testing standards an attractive tool for improving covariate balance in randomized experiments. Despite the intuitiveness of such practice and its arguably already widespread use in reality, the existing literature knows little about its implications on subsequent inference, subjecting many effectively rerandomized experiments to possibly inefficient analyses. To fill this gap, we examine a variety of potentially useful schemes for rerandomization based on p-values (ReP) from covariate balance tests, and demonstrate their impact on subsequent inference. Specifically, we focus on three estimators of the average treatment effect from the unadjusted, additive, and fully interacted linear regressions of the outcome on treatment, respectively, and derive their respective asymptotic sampling properties under ReP. The main findings are twofold. First, the estimator from the fully interacted regression is asymptotically the most efficient under all ReP schemes examined, and permits convenient regression-assisted inference identical to that under complete randomization. Second, ReP improves not only covariate balance but also the efficiency of the estimators from the unadjusted and additive regressions asymptotically. The standard regression analysis, in consequence, is still valid but can be overly conservative. Importantly, our theory is design-based, and holds regardless of how well the models involved in both rerandomization and analysis represent the true data-generating processes.

Professor *Wang Miao* (Peking University) ended the workshop with a talk titled "A stableness of resistance model for nonresponse adjustment with callback data." His abstract is below:

We propose a stableness of resistance assumption for nonresponse adjustment with callback data—a traditional form of paradata that are available in almost all modern surveys to track the data collection process. We establish the identification and the semiparametric efficiency theory without imposing any parametric restrictions, and propose a suite of semiparametric estimation methods including doubly robust ones, which generalize existing parametric approaches for using callback data. We also consider an extension of this framework to causal inference for unmeasured confounding adjustment. Application to a Consumer Expenditure Survey dataset suggests an association between nonresponse and high housing expenditures, and reanalysis of Card (1995)'s dataset on the return to schooling shows a smaller effect of education in the overall population than in the respondents.

4 Scientific Progress Made and Outcome of the Meeting

As intended, the five-day BIRS workshop facilitated fruitful discussions among people from different backgrounds and stimulated more novel ideas about the intersection of disciplines. Our workshop focused on big-data problems related to missingness, mismeasurement, and high dimensionality. In the era of big data, it is critical and timely to bring statisticians, data and computer scientists, and practitioners together to share current research advances in handling novel analytical challenges. This allows us to take a significant step forward in using big data to answer critical scientific questions.