New England Statistical Society's NextGen: Data Science Day 2018

Keynote Speaker and Poster Abstracts

Keynote Address I (9:30AM–10:30AM)
Katherine K. Wallman, Chief Statistician of the United States (1992–2017)

At the Intersection of Evidence and Public Policy: Official Statistics

Daniel Patrick Moynihan, in his eloquent and pointed fashion, advised, "Everyone is entitled to his own opinion, but not to his own facts." Our democracy and economy demand that public and private leaders have unbiased, relevant, accurate, and timely information on which to base their decisions. Taken together, official statistics on demographic, economic, and social conditions and trends are essential to inform decisions that are made by virtually every organization and household. If anything, the Nation's official statistics are at the center of "Evidence-Based Policymaking"; statistics are at the heart of evidence—in a sense these statistics are the "old wine in new bottles." But the system that produces our Nation's official statistics—comprising agencies across every cabinet department and most independent agencies of the Federal Government—is facing growing challenges and opportunities. These challenges and opportunities need to be embraced, even as official statisticians remain true to the principles that have long-guided their work: relevance and timeliness; credibility and accuracy; objectivity; and assurance of confidentiality. While the guiding principles are timeless, official statisticians are challenged to leverage the world around them to meet information needs in new and perhaps better ways. Only then will we stand a chance of presenting the trusted facts that everyone can discuss and debate, and then no doubt agree to disagree, on policy options.

Keynote Address II (3:50PM–4:50PM)
Xiao-Li Meng, Whipple V. N. Jones Professor of Statistics at Harvard University
Founding Editor-in-Chief of Harvard Data Science Review

How Small Are Our Big Data: Turning the 2016 Surprise into a 2020 Vision

The phrase "Big Data" has greatly raised expectations of what we can learn about ourselves and the world in which we live or will live. It also appears to have boosted general trust in empirical findings, because it seems to be common sense that the more data, the more reliable are our results. Unfortunately, this commonsense conception can be falsified mathematically even for methods such as the time-honored ordinary least squares regressions (Meng and Xie, 2014, Econometric Reviews 33: 218-250). Furthermore, whereas the size of data is a common indicator of the amount of information, what matters far more is the quality of data. A 5-element Euler-formula like identity reveals that trading quantity for quality in population statistical inference is a mathematically demonstrably doomed game (Meng, 2018, Annals of Applied Statistics, 685-726). Without considering data quality, Big Data can do more harm than good because of the drastically inflated precision assessment, and hence the gross overconfidence, setting us up to be caught by surprise when the reality unfolds, as we all experienced during the 2016 US presidential election. Data from Cooperative Congressional Election Study (CCES, conducted by Stephen Ansolabehere, Douglas River and others, and analyzed by Shiro Kuriwaki), are used to assess the data quality in 2016 US election polls, with the aim to gain a clearer vision for the 2020 election and beyond.

Both articles are available at https://statistics.fas.harvard.edu/people/xiao-li-meng; the first one is inside Xiao-Li's CV.

New England Statistical Society's NextGen: Data Science Day 2018

Mass Mutual Poster Session Abstracts (4:50pm–7:00pm)

1. Chantal D. Larose, Professional, Eastern Connecticut State University

Identifying Students at Academic Risk via CART Models and Misclassification Costs

Eastern Connecticut State University's Math Foundations Program includes courses which share syllabi, final exams, and required time in the University's math tutoring center, the Mathematics Achievement Center (MAC). These courses are typically taken during a student's first year of college. To aide Eastern's mission of helping students complete their first year of college successfully, we present an analysis of the correlations between MAC participation and final course grades. We use segmented CART models to treat differently the students do not attend the MAC, who attend less than the required time, and who meet or exceed the required time by midterms. The use of misclassification costs which treat false positives as worse than false negatives allow us to focus on students who may fail these courses. By sacrificing accuracy for higher specificity, the model does a better job of finding students who may struggle to successfully complete their first year of college.

2. Jorge Luis Bazan, Professional, University of Connecticut

Performance of asymmetric links and correction methods for imbalanced data in binary regression

In binary regression, imbalanced data result from the presence of values equal to zero (or one) in a proportion that is significantly greater than the corresponding real values of one (or zero). In this work, we evaluate two methods developed to deal with imbalanced data and compare them to the use of asymmetric links. The results based on simulation study show, that correction methods do not adequately correct bias in the estimation of regression coefficients and that the models with power links and reverse power considered produce better results for certain types of imbalanced data. Additionally, we present an application for imbalanced data, identifying the best model among the various ones proposed. The parameters are estimated using a Bayesian approach, considering the Hamiltonian Monte-Carlo method, utilizing the No-U-Turn Sampler algorithm and the comparisons of models were developed using different criteria for model comparison, predictive evaluation and quantile residuals.

3. Shane J. Sacco, Graduate, University of Connecticut

Using Simple Geometry to Quantify Moderation Test Magnitudes

Testing three variable models, particularly moderation models, is increasingly common within public health research. Moderation assesses the degree in which a third variable, the moderator, changes the relationship between the predictor and outcome. While the traditional approach of moderated multiple regression and simple slopes is frequently utilized, it lacks an actual measure of moderation magnitude. Currently, moderation is considered present if the p-value of the predictor-moderator interaction is significant within the moderated multiple regression of the outcome; as the p-value is not an actual effect size, there is no directly quantifiable measure of moderation magnitude. As a novel solution, the present study explored the use of basic geometry to develop a measure of moderation magnitude. With this, predictor-outcome regression lines at high and low values of the moderator were repurposed to create a measurable area within Euclidean space. Parametric behaviors were qualitatively evaluated via simulating two ideal moderation models. Preliminary results indicate that the present area statistic may be the first valid and interpretable measure of moderation magnitude.

4. Eduardo Schneider Bueno de Oliveira, Graduate, University of Connecticut

An Application of DINA model to Beck Depression Inventory data

Cognitive diagnosis models (CDMs) are useful psychometric tools which allows to identify test-respondents' profile concerning to a set of latent attributes underlying a latent variable, such as a cognitive skill, a psychological trait, or an attitude. The deterministic inputs, noisy "and" gate (DINA) model is one of the most featured models in the CDMs literature, because of its parsimony and easiness of interpretation. The DINA model considers a test with J dichotomous items and N respondents to classify them based upon their mastery and non-mastery of K dimensions previously set. In this work we analyze data from the University of São Paulo, Brazil, containing 1,111 college students' responses to the Beck Depression Inventory. We propose the use of DINA model, through a Bayesian approach, analyzing the behavioral pattern concerning to depression and the items performance. Our intention is to show the possibility of applying DINA model and its benefits to the analysis of such kind of data, motivating more in-depth studies and discussions in interdisciplinary researches involving not only statisticians but, as well, specialists in psychiatry and other scientific fields.

5. Devin J. McConnell, Undergraduate, University of Rhode Island

Malaria Genome Population Identification With HMM Signals

There is a global effort to eradicate malaria caused by several species of the genus Plasmodium. To that end, the Pf3k project has collected 3000 samples of P. falciparum from 14 different countries. To understand the effects of eradication, analysis must be done to learn about the populations' structures. Our goal is to be able to identify, with genomic data, the parasitic population a particular infection stems from. We hypothesize that for a majority of the data there would be a single population per country. We used Mauve to build consensus sequences and HMMER to build profile hidden Markov models for each alignment. Using 10-fold cross-validation, we visualized the true positive rate against the false positive rate by testing the models and plotting ROC curves. This method provided results of $AUC \geq 0.7$ in some countries, representing models that successfully identify the population of malaria genomes. Based on our cross-validation results, we have sub-divided several populations into region-subregion pairs. In future work, we will provide hidden Markov models for each subpopulation, providing a valuable tool to identify non-native malaria infections within a geographic region.

6. Jai Woo Lee, Graduate, Dartmouth College

Penalized Estimation of Sparse Concentration Matrices Based on Prior Knowledge with Applications to Placenta Elemental Data

Identifying patterns of association or dependency among high-dimensional biological datasets remains a challenge. Thus, analyzing sparse precision matrices which define interactions of elements in the datasets is essential. In this paper, we introduce a weighted sparse Gaussian graphical model that can incorporate prior knowledge to infer the structure of the network of trace element concentrations, including both essential elements and toxic metals present in the human placental biopsies from the New Hampshire Birth Cohort Study. The chemical architecture for elements is complex; hence, the proposed method was applied to infer the dependency structures of the elements using prior knowledge of their biological roles. Results demonstrate that weighting elements which had a high number of neighbors in the network significantly increased accuracy in estimating interactions of elements. Our method also successfully identified fundamental elemental associations consistent with known chemical and biological roles which contained a separate network of Ca and K, an interconnected sub-network of Mg, P, Ba and Sr, and the appearance of Zn as a hub of multiple elemental associations.

7. Bruna Folegatti Santana, Graduate, University of Connecticut

Association between runs of homozygosity pattern and low variability of progeny's estimated breeding values in Nellore cattle

The purpose of this work is to identify patterns of runs of homozygosity (ROH) in a group of bulls with low variability of their progeny's estimated breeding values (EBVs). Seventy-nine genotyped Nellore bulls were selected from a previous existing database. Each one of those had at least 20 progeny with previously calculated EBVs. The EBVs were estimated by fitting univariate (linear) animal models, using the methodology of restricted maximum likelihood. The traits analyzed included weight at 18 months, precocity and muscularity. ROH were identified, and two groups were stablished, the first with bulls that presented low variability of their progeny's EBVs (LV group) and the second one with high variability (HV group). Bulls from the LV group presented higher proportion of ROH than HV bulls, when ROH minimum size was 16Mb. Correlations between the standard deviation of progeny's EBVs and number of ROH were negative, and stronger in the LV groups, especially when the minimum segment length was 1, 2, 4, 8 and 16 Mb. Results from this study suggests an association between presence of ROH and lower variability of progeny's breeding value. hidden Markov models for each subpopulation, providing a valuable tool to identify non-native malaria infections within a geographic region.

8. Seokchae Hwang, Graduate, Syracuse University

Optimal Policies under existing Heterogeneous Treatment effects of Job Training Programs

Estimating heterogeneous treatment effects (HTE) have not developed much yet in economic literature. Thanks to recent development of machine learning methods and rich administrative data, we estimate HTE of Korean Job Training Programs (JTPs). Despite significant positive average treatment effects on employment, our estimation results using "honest causal forest" suggest that substantial amount of individuals are affected negatively. Specifically, around 25% of unemployed workers are worsen by taking JTPs. Having this information, we define endogenous quartiles of HTE. We present average characteristics of those groups and find moderate differences between the first quartile (with significant negative effects) and fourth quartile (the most favorable group). Also, our results suggest an optimal stepwise decision process for avoiding negative effects, so that can improve policy efficiency as well as unemployed workers' welfare.

9. Bo Ning, Professional, Yale University

Bayesian methods for high-dimensional data analysis

Analyzing large and complex datasets is one of the most challenging tasks for modern statisticians. The challenging comes from when the model has the number of parameters is bigger than the number of observations, so that classical methods cannot be used to carry out the analysis. Statistical inference is available if one is willing to assume the true parameter is sparse. Recently, several Bayesian methods have been developed for analyzing these datasets based on this assumption. I will present two examples to demonstrate the benefit of using these methods—one analyzes a complex astronomy dataset, and the other studies the sales lift of an advertising campaign for multiple brick-and-mortar stores. I will also provide several theoretical justifications for using these methods, including deriving the posterior contraction rate and conditions for selection consistency, and quantifying uncertainties for the parameter of interest.