Newsworthy Research Highlights from JSM 2025

The 2025 Joint Statistical Meetings will bring together statisticians and data scientists from around the world from Saturday, August 2, to Thursday, August 7. This year, JSM will be held in Nashville, Tennessee. This tip sheet highlights interesting presentations from the conference. Complimentary press registration is open, courtesy of the ASA. Email edoffice@amstat.org for more information.

Featured Research (Synopses Below)

Monday Highlights

1. Data Privacy and Statistical Science
2. Syntax-Guided Diffusion Large Language Model for Personalized Text Generation
3. Advancing Evidence for Opioid Use Disorder Treatments Through Real-World Data and Novel Statistical Methods

Tuesday Highlights

4. What Kind of Music Do You Like? A Statistical Analysis of Music Genre Popularity Over Time
5. Can We Trust LLM Without High Confidence?
6. What Is Analytic Fluency? A Thematic Content Analysis of Interviews with Expert Data Analysts

Wednesday Highlights

7. Expected Points Above Average: A Novel NBA Player Metric Based on Bayesian Hierarchical Modeling
8. CancerLLM: A Large Language Model in Cancer Domain
9. Rethinking Suicide Prevention Research: Moving Beyond Traditional Statistical Significance

Thursday Highlights

10. Updating Draft Value Pick Charts

11. Applications of AI-Generated Digital Twins to Improve Efficiency and Decision-Making in Clinical Trials
12. Leveraging AI Protein Structure Prediction for Functional Annotation of the Human Virome

Synopses

Monday

1. **Data Privacy and Statistical Science**

Social media, AI, and emergency evacuations are just a few examples of how personal information can enhance our lives and contribute to the public good. However, this use of often confidential data comes with the potential risk of compromising personal privacy. In this Introductory Overview Lecture, we discuss how statistical science concepts and methods can be applied to protect the confidentiality of sensitive data while maintaining its value for social good. We review different statistical approaches to measuring confidentiality risks and utility in data releases, as well as recent advances for reducing those risks.

2. **Syntax-Guided Diffusion Large Language Model for Personalized Text Generation**

Large language models have demonstrated success in generating human-like text. However, sentences generated by LLMs (e.g., ChatGPT) tend to be generic and lacking personalized characteristics. Recent development on diffusion models has shown the potential in diversified generation and iterative refinement; however, its limitations still exist, especially when the generated text is complicated. This presentation will discuss a syntax-guided diffusion model to achieve both well-written and personalized text generation.

3. **Advancing Evidence for Opioid Use Disorder Treatments Through Real-World Data and Novel Statistical Methods**

The opioid epidemic remains a major public health crisis. Although evidence-based treatments for opioid use disorder exist, most people with OUD do not receive treatment. Pragmatic trial designs have therefore been proposed to evaluate interventions designed to increase OUD treatment within entire clinics or health systems by leveraging health records and other real-world data sources. In this talk, we present case studies that illustrate key challenges of using real-world data for evaluating intervention effects and highlight novel statistical methods that have been

developed and are being implemented in these case studies to provide robust evidence on intervention effects to improve outcomes of people with OUD.

Tuesday

4. **What Kind of Music Do You Like? A Statistical Analysis of Music Genre Popularity Over Time**

Popular music genre preferences can be measured by consumer sales, listening habits, and critics' opinions. We analyze trends in genre preferences from 1974 through 2018 presented in annual Billboard Hot 100 charts and annual Village Voice Pazz & Jop critics' polls. We model yearly counts of appearances in these lists for eight music genres with two multinomial logit models, using various demographic, social, and industry variables as predictors. Since the counts are correlated over time, we use a partial likelihood approach to fit the models. Our models provide strong fits to the observed genre proportions and illuminate trends in the popularity of genres over the sampled years, such as the rise of country music and the decline of rock music in consumer preferences and the rise of rap/hip-hop in popularity among both consumers and critics.

5. **Can We Trust LLM Without High Confidence?**

As large language models expand their presence across diverse applications, gauging and understanding their inherent uncertainty is crucial. Estimating uncertainty yields deep insights into the reliability of LLM predictions, playing a pivotal role in decision-making processes within high-stakes domains like medical diagnostics and robotic emergency response, where erroneous outputs may lead to significant adverse effects. Our goal is to encourage the advancement of more dependable and efficient approaches to expressing, estimating, and calibrating uncertainty in LLMs for real-world applications.

6. **What Is Analytic Fluency? A Thematic Content Analysis of Interviews with Expert Data Analysts**

It is common sense that data should be analyzed well, rather than badly. Despite this, the actual criteria by which we judge the quality of an analysis are opaque, intuitive, and heavily influenced by the uncertain standards of disciplinary norms, routines, or subjective judgments of what "feels right" or "seems off." This lack of explicit criteria is problematic not just for analysts facing real challenges in their work, but also for hiring, program evaluation, and teaching. Indeed, many analysts report their training left them unprepared for the challenges faced in real-world

analytic settings. To better understand what good analysis looks like, we conducted a qualitative study using grounded theory methodology in a sample of highly experienced analysts from diverse professional backgrounds. Our aim was to more explicitly identify the content of what we call analytic fluency, or the "soft skills" of data analysis used in real-world settings. Our analysis uncovered five rich, higher-order themes (i.e., families of skills) with 11 lower-order sub-themes. We present these findings and consider their implications for data analysis practice.

Wednesday

7. Expected Points Above Average: A Novel NBA Player Metric Based on Bayesian Hierarchical Modeling

Team and player evaluation in professional sport is extremely important given the financial implications of success/failure. It is especially critical to identify and retain elite shooters in the National Basketball Association, one of the premier basketball leagues worldwide, because the ultimate goal of the game is to score more points than one's opponent. To this end, we propose two novel basketball metrics: "expected points" for team-based comparisons and "expected points above average" as a player-evaluation tool. Both metrics leverage posterior samples from Bayesian hierarchical modeling framework to cluster teams and players based on their shooting propensities and abilities. We illustrate the concepts for the top 100 shot takers over the last decade and offer our metric as an additional metric for evaluating players.

8. CancerLLM: A Large Language Model in Cancer Domain

This talk will introduce large language models were trained on local health care data to conduct multiple downstream tasks such as phenotyping extraction, diagnosis generation, and treatment suggestions. The CancerLLM outperformed other state-of-the-art models on these cancer domain-specific tasks. The CancerLLM can also be used for disease prognosis tasks such as cardiotoxicity prediction models.

9. The Data Science of Refugee Camp Management

The confluence of climate crises and population growth indicate the world will see more refugee camps in the future. We need to get smarter about how to manage such camps, and one component of camp management rests squarely on traditional official statistics. One needs to know the kinds of data a census bureau collects and the kinds of data a public health service gathers. One needs to track education and camp

security. And it is vital to improve the on-boarding and off-boarding processes in a refugee camp. This talk describes preliminary work in that direction.

Thursday

10. Updating Draft Value Pick Charts

Draft pick value charts are tools professional sports teams use to give currency to draft picks. Draft picks are opportunities to select new players entering their leagues. There is a modest body of work in this area that considers the performance of players in their first several years post-draft or their second contract values as responses. These outcomes require a sometimes lengthy delay between the original draft picks and observation of the response. This delay can be problematic if player performance distributions or team behaviors are fluctuating over time. In this paper, we propose a methodology for updating a draft value pick chart by imputing future player performance. This imputation is done via a flexible generalized additive model accounting for drafted player performance to date and a player's draft selection. We apply this methodology to data from the National Hockey League and National Football League.

11. Applications of AI-Generated Digital Twins to Improve Efficiency and Decision-Making in Clinical Trials

Effective and rapid decision-making in clinical trials requires unbiased and precise treatment effect inferences. Recent advances in artificial intelligence are poised to revolutionize breakthroughs in innovative trial design and analyses, improving efficiency in Phase 2 and 3. We present AI-enabled methods that combine digital twins and traditional statistical frameworks to improve trial efficiency that satisfy regulatory guidance. Digital twin generators are pre-trained generative models that generate digital twins for each trial participant using only baseline measurements and are fully pre-specifiable. We present results of recent case studies and discuss prospective applications of these methodologies.

12. Leveraging AI Protein Structure Prediction for Functional Annotation of the Human Virome

The study of the human virome is advancing rapidly, adding a crucial layer to understanding human health and disease mechanisms. Current virome analysis unit, vOTU, relies solely on nucleotide sequence similarity, which is limited by virus mutation rates and lacks functional insights, leading to unreliable and unreproducible results. We propose clustering viruses based on their functions, which are determined

by the 3D structures of the proteins they produce. We will leverage AlphaFold2, a cutting-edge AI system, to accurately and massively predict the protein 3D structures based on the amino acid sequences translated from viral genomes, and then develop a scalable, accurate pipeline for clustering viral proteins and taxa, resulting in a function-based virus catalog. Compared to the existing clustering of metagenomic gut virus catalogue, our algorithm demonstrates more robust and informative clustering.