Newsworthy Research Highlights from JSM 2024

The 2024 Joint Statistical Meetings will bring together statisticians and data scientists from around the world from Saturday, August 3, to Thursday, August 8. This year, JSM will be held in Portland, Oregon. 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. Should Time Duration Be Considered in Assessing Vaccine Performance? Why and How?
2. National Experimental Well-Being Statistics
3. Working at the Intersection of Statistics and AI Policy

Tuesday Highlights

4. Detecting and Mitigating Algorithmic Bias in Online Misinformation
5. Uncertainty Quantification for Low-Likelihood High-Impact Weather Events Using Spatio-Temporal Statistical Modeling
6. Fostering Inclusiveness in Big Data: Changing the Normative Process for Selecting Covariates in Statistical Models

Wednesday Highlights

7. Bias and Fairness in Generative AI
8. Evaluating the (Moral) Beliefs Encoded in LLMs
9. Rethinking Suicide Prevention Research: Moving Beyond Traditional Statistical Significance

Thursday Highlights

10. Tracking U.S. Consumers in Real Time with a New Weekly Index of Retail Trade
11. Quantifying Misdiagnosis-Related Harm Leveraging Health Record Data and Through Mixture-Model–Based Novel Measures
12. A Spatial Statistical Framework for Detection and Attribution of Extreme Precipitation

Synopses

Monday

1. Should Time Duration Be Considered in Assessing Vaccine Performance? Why and How?

Current procedure requires the vaccine efficacy, an easy-to-understand overall index of vaccine performance, be above certain level such as 50% in addition to certain safety requirements. However, the vaccine efficacy does not take time duration into consideration and only presents an overall performance assessment. It may be severely misleading, as we show in a simple example. We demonstrate vaccine performance can be assessed with a novel simple statistical quantity that takes both protection performance and time duration into consideration. This method has several desirable statistical properties and is also easy to understand and compute.

We will illustrate this method with three real data sets from the initial clinical trials of COVID-19 vaccine studies submitted to the FDA for emergency use.

2. National Experimental Well-Being Statistics
We summarize the National Experimental Wellbeing Statistics (NEWS) Project, which makes three unique contributions. First, it simultaneously addresses many sources of measurement error, including unit and item nonresponse and underreporting in surveys, as well as the various challenges in administrative data such as measurement error, conceptual misalignment, and incomplete coverage. Second, it brings together all the available survey and administrative data, which allows us to address many of the shortcomings of individual data sources. Third, it proposes a model to combine survey and administrative earnings data given measurement error in both sources.

3. Working at the Intersection of Statistics and AI Policy
Statisticians play critical roles in the design, implementation, and evaluation of AI policy. Roles for statisticians include developing assessments of societal impacts of AI, including workforce impacts, economic benefits, and costs; surveys of public attitudes; evaluations of policy frameworks, guidelines, and standards; and methods of measuring effects on innovation rates. To accomplish fairness metrics and bias detection; scalable methods of model evaluation and benchmarking, statisticians are critical contributors. In addition, security measures such as robustness to adversarial attacks and data poisoning; privacy-preserving AI techniques such as differential privacy, secure multi-party computation, and federated learning; and comparisons of AI capabilities across countries all require sound statistical methodology.

Tuesday
4. Detecting and Mitigating Algorithmic Bias in Online Misinformation
There is perhaps no bigger issue facing our world right now than misinformation. The advent of tools like ChatGPT has increased this risk. A central reason for this is bias from large language models and how that can lead to misleading and/or incorrect information disproportionately impacting certain communities. NORC is developing a model of online information to better understand how to detect and mitigate bias in such models. The data are focused specifically on the topic of COVID vaccine misinformation, which the study team chose because of the strong historical record of misinformation across social media platforms and issues related to health equity. NORC collected more than 10 terabytes of data from across Twitter and Instagram from 2020 to 2023. The study team hand-coded a training sample, building upon several open-source misinformation indexes, and then trained and deployed the model. This presentation will share learnings from this process, share the model developed, and finally describe the learnings gleaned from the process related to bias in LLM development.

5. Uncertainty Quantification for Low-Likelihood High-Impact Weather Events Using Spatio-Temporal Statistical Modeling
Determining the probability and severity of a low-likelihood high-impact weather event from the historical record is difficult due to their relatively rare occurrence. Instead, we shift our focus to the drivers of the climatology surrounding weather events. Specifically, we represent the climate system as the sum of two parts—the climatological forcing and internal variability—and model the drivers of these two processes. We model the climatological forcing as changes in the system

due to anthropogenic-induced climate change using a set of measurable variables. The internal variability represents the variation in the system due to its natural cycle, which we model using Bayesian singular value decomposition, where the basis functions in the decomposition capture the spatial and temporal modes of variability. By decomposing the climate system in terms of its climate forcing and internal variability, we can determine which combination of the drivers result in high-impact weather events and the probability of these events occurring. We apply our framework to two-meter air temperature in the Pacific Northwest, providing additional insight into the 2021 heatwave.

6. Fostering Inclusiveness in Big Data: Changing the Normative Process for Selecting Covariates in Statistical Models

In observational and/or epidemiological research, we often adjust for a host of socio-demographic variables to clarify the influence of a predictor on an outcome. However, we rarely use all possible covariates. Instead, we use what is available or the normative variables we have been taught to include. If we want to be more inclusive of underrepresented communities, we must start making the invisible visible by showing representation in our statistical analyses. This presentation addresses ways to be more inclusive in the ways we select and use covariates in statistical analyses.

Wednesday

7. Bias and Fairness in Generative AI

Recent innovations in generative AI imply multimodal systems that can generate and analyze across a variety of data formats—including texts, images, and graphs—find usage across various aspects of modern life, including health care, transportation, climate, and finance. However, such rapid advances also give rise to ethical and governance challenges that must be addressed. In this work, we outline ethical challenges across various modes of generation and identify ways to detect the extent of the challenges. In addition, we outline statistical/machine learning strategies to mitigate some of the concerns outlined.

8. Evaluating the (Moral) Beliefs Encoded in LLMs

This talk concerns the design, administration, post-processing, and evaluation of surveys on LLMs. It has two parts: 1. A statistical method for eliciting beliefs encoded in LLMs. We introduce statistical measures and evaluation metrics that quantify the probability of an LLM "making a choice," the associated uncertainty, and the consistency of that choice. 2. We apply this method to study moral beliefs encoded in LLMs, especially in ambiguous cases where the right choice is not obvious. We design a survey with high-ambiguity moral scenarios (e.g., "Should I tell a white lie?") and low-ambiguity scenarios (e.g., "Should I kill an innocent person?"). Each scenario includes a description, two possible actions, and labels indicating violated rules (e.g., "Do not kill."). We survey 28 open- and closed-source LLMs. We find (a) most LLMs choose actions that align with human annotators in unambiguous scenarios but reflect uncertainty in ambiguous situations, (b) some LLMs reflect high uncertainty because their responses are sensitive to the question wording, and (c) some models reflect clear preferences in ambiguous scenarios. Specifically, closed-source models agree with each other.

9. Rethinking Suicide Prevention Research: Moving Beyond Traditional Statistical Significance

Suicide is a major public health concern, and despite decades of research, there has been a disappointing lack of progress. We argue the over-reliance on traditional statistical significance cutoffs and the failure to consider marginal findings are limiting the clinical benefits of research. Many reviewers and editors still insist on "statistically significant" results at $p<0.05$ for publication. Thus, many potentially promising results receive less visibility among clinicians making treatment decisions. The American Statistical Association has called upon researchers to view the *p*-value as continuous, with the call being adopted by leading journals, including the *New England Journal of Medicine* and *Psychological Science*. However, most suicide journals do not explicitly discourage or require *p*-values and best practices. We want to call upon suicide researchers to be more open to marginal findings that suggest promising trends for suicide prevention strategies and interventions.

Thursday

10. Tracking U.S. Consumers in Real Time with a New Weekly Index of Retail Trade

The Chicago Fed Advance Retail Trade Summary (CARTS) uses high-frequency data on payment card transactions, retail foot traffic, gas sales, consumer sentiment, and online prices to nowcast the U.S. Census Bureau's Advance Monthly Retail Trade Summary. We show how the resulting weekly indexes of retail trade and prices can be used to improve upon real-time predictions for both retail sales and prices.

11. Quantifying Misdiagnosis-Related Harm Leveraging Health Record Data and Through Mixture-Model–Based Novel Measures

Investigating and monitoring misdiagnosis-related harm is crucial for improving health care. However, this effort has traditionally focused on the chart review process, which is labor intensive, potentially unstable, and does not scale well. To monitor medical institutes' diagnostic performance and identify areas for improvement in a timely fashion, researchers proposed to leverage the relationship between symptoms and diseases based on electronic health records or claim data. Specifically, the elevated disease risk following a false-negative diagnosis can be used to signal potential harm. We proposed a mixture regression model and related harm measures and profiling analysis procedures to quantify, evaluate, and compare misdiagnosis-related harm across institutes with potentially different patient population compositions. We studied the performance of the proposed methods through simulation studies. We then illustrated the methods through data analyses on stroke occurrence data from the Taiwan Longitudinal Health Insurance Database. From the analyses, we quantitatively evaluated risk factors for being harmed due to misdiagnosis, which unveiled some insights for health care quality.

12. A Spatial Statistical Framework for Detection and Attribution of Extreme Precipitation

In climate sciences, great effort is taken to determine whether reported changes in the distribution of climate variables are due to human influences or anthropogenic forcings. To answer this question, it is common to use climate data—typically the output of climate models obtained under different forcings—and see if there is a significant association between the observed changes in a given climate variable and the climate model output. Recently, these types of studies have been performed using observational data rather than climate model output data. While this approach offers some advantages, it introduces uneven spatial distribution and potentially the preferential sampling nature of the data.

In this talk, first using simulations we investigate whether the unequal spatial distribution of the data does lead to issues and erroneous conclusions in terms of detection and attribution statements, then we present a spatial statistical framework that allows us to determine whether local changes in extreme precipitation are due to human activity while accounting for preferential sampling.