What are the Most Important Statistical Ideas of the Past 50 Years?

Faryal Fodderwala

November 18, 2024

Introduction

- Overview of 8 significant statistical ideas from 1970 to 2021.
- Authors: Andrew Gelman and Aki Vehtari.
- Purpose: To provoke thought and discussion about modern statistical innovations and their impact on data science.

Authors' Background

- **Andrew Gelman**:
  - Professor of Statistics and Political Science, Columbia University.
  - Renowned for Bayesian statistics and multilevel modeling.
- **Aki Vehtari**:
  - Professor of Computational Probabilistic Modeling, Aalto University.
  - Focused on Bayesian computation and model assessment.

Overview of the Paper

- Timeframe: 1970 to 2021, focusing on the development of modern statistics.
- 8 statistical ideas selected based on their influence on statistical theory, computation, and applications.
- Emphasis on integrating computation with statistical modeling.

Counterfactual Causal Inference

- Allows causal inference using observational data.
- Framework based on "potential outcomes" or "counterfactuals."
- Example: Studying the effect of NYC's "Vision Zero" traffic policy using observational data.

$$\text{Causal Effect: } Y(1) - Y(0)$$

- $Y(1)$: Outcome if treated.
- $Y(0)$: Outcome if untreated.
- Challenge: Only one outcome is observed.

Real-World Connection

NYC Open Data provides datasets on traffic accidents, enabling causal analysis of interventions like "Vision Zero."

Bootstrapping and Simulation-Based Inference

- Introduced by Bradley Efron (1979).
- Resampling technique to estimate sampling distributions without assumptions about data distribution.

Algorithm:

1. Resample the dataset with replacement.
2. Compute the statistic of interest (e.g., mean).
3. Repeat n times to estimate variability.

Example: NYC 311 Calls Data

Use bootstrapping to estimate variability in the average response time for complaints across boroughs.

Overparameterized Models and Regularization

- High-dimensional models with more parameters than data points.
- Regularization prevents overfitting by adding penalties to the model:

$$\text{LASSO: } \min \left(||Y - X\beta||^2 + \lambda ||\beta||_1 \right)$$

Example

Neural networks for NYC Open Data crime prediction:

- Regularization reduces noise and ensures generalizable predictions.

Bayesian Multilevel Models

- Models hierarchical data with varying parameters at different levels.
- Example: Modeling housing prices across NYC boroughs.

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \epsilon_{ij}$$

- u_j: Random effect for borough j.
- ϵ_{ij}: Error term for observation i in borough j.

Advantage

Combines individual-level and group-level variability for improved estimates.

Generic Computation Algorithms

- Advances in algorithms like MCMC, EM, and variational inference.
- Enabled complex models and large-scale Bayesian analysis.

Connection to NYC Open Data

Use MCMC to model traffic flow patterns and predict congestion hotspots.

Adaptive Decision Analysis

- Framework for making decisions during experiments.
- Application: Stopping clinical trials early for ethical reasons.

Real-World Example

In NYC public health studies, adaptive analysis helps evaluate the success of vaccination campaigns.

Robust Inference

- Focuses on reliability under model misspecification.
- Example: Median-based estimators for income disparity in NYC.

Key Insight

Robust inference allows valid results even when data deviates from assumptions.

Exploratory Data Analysis (EDA)

- Emphasizes visualization and insights over strict models.
- Examples: Trends in NYC Open Data on crime or health disparities.

Connection to NYC Open Data

- Apply statistical methods to NYC datasets.
- Example: Visualize and analyze health disparities using robust inference and EDA.

Conclusions and Future Directions

- These statistical ideas are foundational to modern data analysis.
- Future: Integration of machine learning with causal inference.
- Importance of robust and interpretable models for real-world applications.

The Importance of Human Oversight in Statistical Innovations

- As computational power advances, machine learning and statistical algorithms can model complex systems.
- However, these models are only as good as the assumptions and data they are based on.
- Example: Self-driving cars can use machine learning to navigate, but human oversight is needed to determine:
  - Are the outcomes (e.g., accident rates) statistically significant?
  - Are the algorithms operating ethically and equitably?
- **Key Point**: Computational tools are powerful, but without human observation and ethical guidance, they can lead to unintended consequences.

Reflection from Gelman

"On one hand, you have all these amazing things that machine learning can do, like self-driving cars, but you'll need a statistician to tell you if the number of people being killed by the self-driving cars is statistically significant." – Paraphrased from Andrew Gelman

Questions?

Thank you! Any questions?