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

Andrew Gelman and Aki Vehtari

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

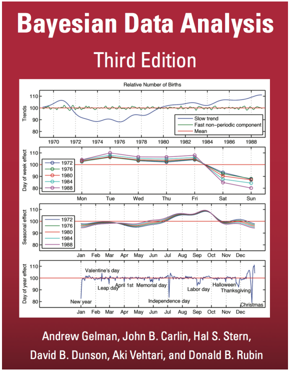
Andrew Gelman

- **Andrew Gelman**:
  - Professor of Statistics and Political Science, Columbia University.
  - Renowned for Bayesian statistics and multilevel modeling.

Aki Vehtari

- **Aki Vehtari**:
  - Professor of Computational Probabilistic Modeling, Aalto University.
  - Focused on Bayesian computation and model assessment.

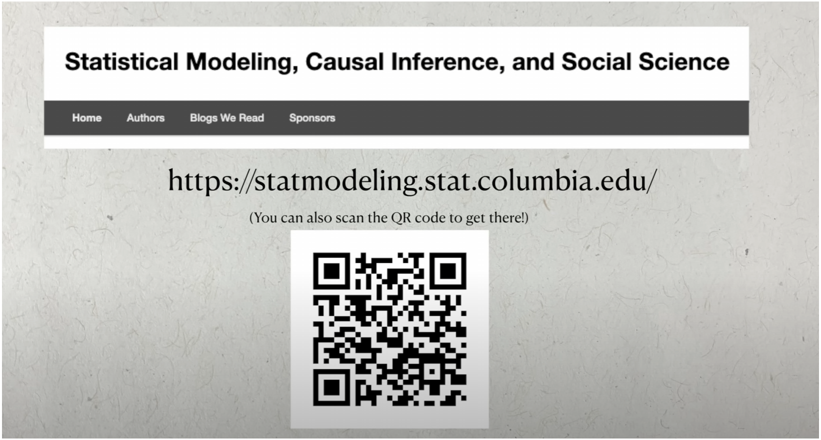
Authors' Background (cont.)



Bayesian Data Analysis

- **The Book: Bayesian Data Analysis**
  - Written by Andrew Gelman, Aki Vehtari, and others.
  - Widely regarded as the foundational text ("the bible") for Bayesian practitioners.
  - Covers theory, computation, and applied Bayesian methods.
- **Their Authority on the Topic**
  - Through this book, Gelman and Vehtari have shaped the modern understanding of Bayesian statistics.
  - Their extensive research and contributions give them unique insights to answer: *What are the most important statistical ideas of the past 50 years?*

Gelman’s Blog



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.
- Designed as an **essay**, not a research manuscript.
- Acknowledges that no definitive list can encompass all significant ideas.

01: Counterfactual Causal Inference

- Allows causal inference using observational data.
- Framework based on "potential outcomes" or "counterfactuals."

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

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

02: 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.

03: 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 with regularization techniques balance flexibility and robustness.

04: Bayesian Multilevel Models

- Models hierarchical data with varying parameters at different levels.
- Example: Aggregating data across different groups in meta-analysis.

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

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

Advantage

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

05: Generic Computation Algorithms

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

06: Adaptive Decision Analysis

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

07: Robust Inference

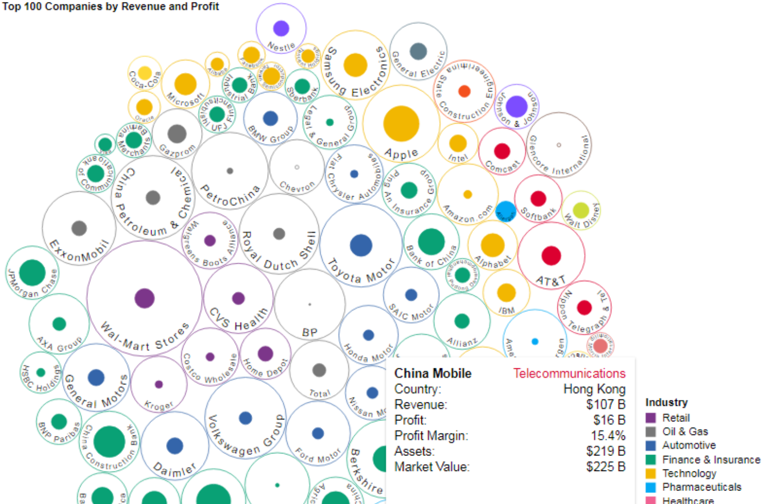
- Focuses on reliability under model misspecification.
- Example: Median-based estimators and propensity score matching.

Key Insight

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

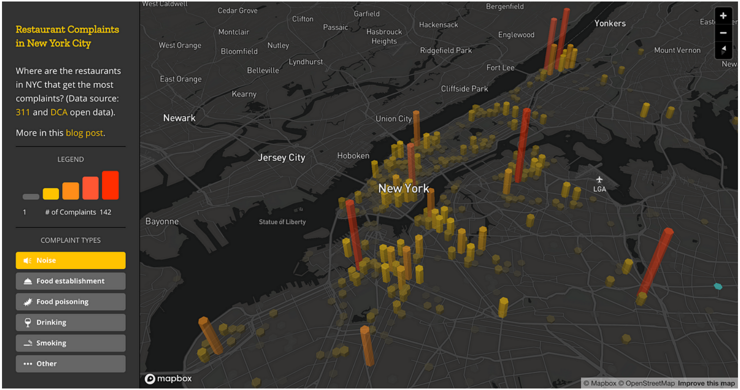
08: Exploratory Data Analysis (EDA)

- Emphasizes visualization and insights over intense theory and computation.
- Useful in understanding the relation between data, fitted model, and predictions.



Connection to NYC Open Data

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



https://labs.mapbox.com/bites/00304/

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?

References

- Gelman, Andrew, and Aki Vehtari. 2021. "What are the Most Important Statistical Ideas of the Past 50 Years?" *Journal of the American Statistical Association* 116, no. 536: 2087–2097.
- Gelman, Andrew. *Statistical Modeling, Causal Inference, and Social Science*. Accessed at: `https://statmodeling.stat.columbia.edu/`
- Gelman, Andrew. *What are the most important statistical ideas of the past 50 years?* YouTube video. Available at: `https://www.youtube.com/watch?v=M6ha2UeSZbo`
- Gelman, Andrew. Additional commentary on statistical ideas. YouTube video. Available at: `https://youtu.be/nCyGhqQWj2g?si=GM9KpuWtg4tGV8je`
- Mapbox. 2018. "Exploring NYC Open Data with 3D Hexbins." Available at: `https://blog.mapbox.com/exploring-nyc-open-data-with-3d-hexbins-5af2b7d8bc46`