

# 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

- ▶ **Andrew Gelman:**

- ▶ Professor of Statistics and Political Science, Columbia University.
- ▶ Renowned for Bayesian statistics and multilevel modeling.

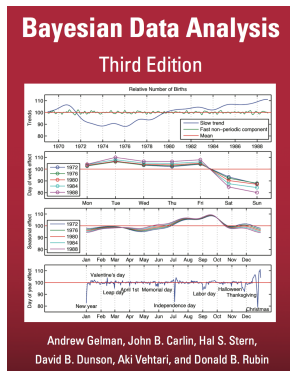


Aki Vehtari

- ▶ **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?*

## Statistical Modeling, Causal Inference, and Social Science

[Home](#)[Authors](#)[Blogs We Read](#)[Sponsors](#)

<https://statmodeling.stat.columbia.edu/>

(You can also scan the QR code to get there!)



# 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.
- ▶ Designed as an **essay**, not a research manuscript.
- ▶ Meant to provoke thought, invite discussion, and reflect on statistical progress.
- ▶ 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.

## Real-World Connection

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

## 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.

### Example: NYC 311 Calls Data

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



## 03: Overparameterized Models and Regularization

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

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

### Example

Neural networks for NYC Open Data crime prediction:

- ▶ Regularization reduces noise and ensures generalizable predictions.

## 04: 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  $j$ .
- ▶  $\epsilon_{ij}$ : Error term for observation  $i$  in borough  $j$ .

### 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.

### Connection to NYC Open Data

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

## 06: 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.

## 07: 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.

## 08: 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?

# 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>