

# 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.



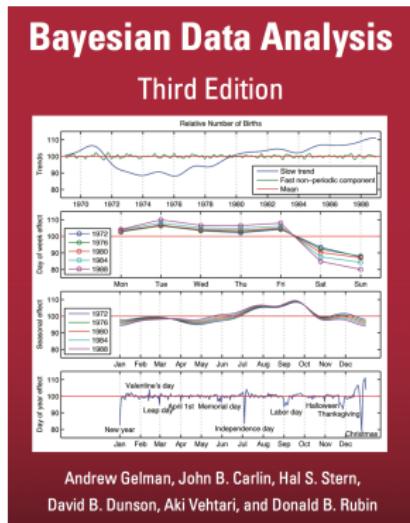
Aki Vehtari

► **Aki Vehtari:**

- ▶ Professor of Computational Probabilistic Modeling, Aalto University.
- ▶ Focused on Bayesian computation and model assessment.

# Authors' Background (cont.)

## ► The Book: Bayesian Data Analysis



*Bayesian Data Analysis*

## ► 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.
- ▶ Published to the *Journal of the American Statistical Association* as an **essay**, not a research manuscript.
- ▶ Acknowledges that no definitive list can encompass all significant ideas.

# Key Statistical Concepts

## The 8 Statistical Ideas Covered in the Paper:

1. Counterfactual Causal Inference
2. Bootstrapping and Simulation-Based Inference
3. Overparameterized Models and Regularization
4. Bayesian Multilevel Models
5. Generic Computation Algorithms
6. Adaptive Decision Analysis
7. Robust Inference
8. Exploratory Data Analysis (EDA)

**Purpose:** These concepts reflect significant innovations in statistical theory, computation, and application from the last 50 years.

# The Challenge of Observational Data

- ▶ **In an ideal world: Experimental Data**
  - ▶ Researchers control interventions.
  - ▶ Participants are assigned to "treatment" and "control" groups.
  - ▶ Enables *causal claims* about the effect of an intervention.
- ▶ **In the real world: Observational Data**
  - ▶ Researchers cannot control who receives the intervention.
  - ▶ Only allows for *correlational claims*.
  - ▶ Several fields of study are prone to having more observational data i.e. statistics, econometrics, psychometrics, epidemiology, and computer science.

**Challenge:** How can we infer causality in the absence of experimental data?

# Introducing Counterfactuals

## ► Model Setup:

- ▶ Let  $X$  represent attending class:
  - ▶  $x = 0$ : Did not attend class.
  - ▶  $x = 1$ : Did attend class.
- ▶ In reality, I choose to attend class and receive a score  $Y_1$  on my paper.

## ► Key Question: Did attending class improve my score?

- ▶ To answer this, we need to consider an alternative reality:
- ▶ What score ( $Y_0$ ) would I have received if I *did not* attend class?

## Key Idea: Counterfactuals

- ▶  $Y_0$ : The score in the unobserved reality (if I did not attend class).
- ▶ This unobserved reality is called the *counterfactual*.

# 01: Counterfactual Causal Inference

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

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

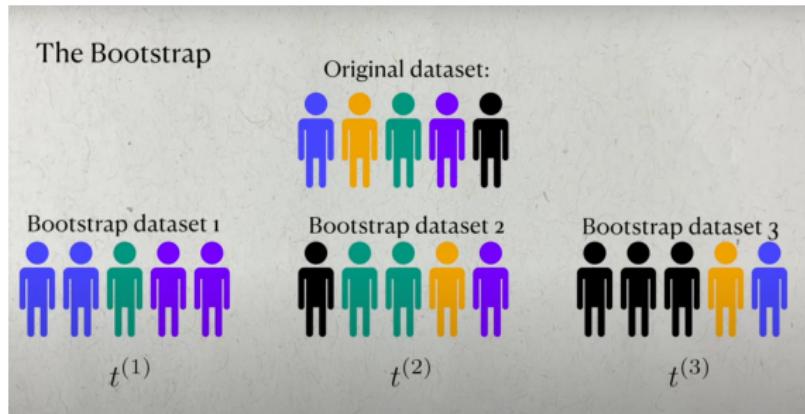
- ▶ Introduced by Bradley Efron (1979).
- ▶ Resampling technique to estimate sampling distributions without assumptions about data distribution.

### Algorithm:

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

## 02: Bootstrapping and Simulation-Based Inference (cont.)

### ► The Bootstrap Algorithm:



- $\hat{F}(t)$  Goal: Estimate the CDF of some statistic.
- **Intuition:** Data sampled from the original dataset resembles a new dataset.
- For each bootstrap dataset, calculate a statistic of interest and derive its distribution:

$$t^{(1)}, t^{(2)}, t^{(3)} \rightarrow \hat{F}(t)$$

# Simulation-Based Inference (cont.)

- ▶ Example using Bayes' Rule:

$$P(\theta | X) = \frac{P(X | \theta) P(\theta)}{P(X)}$$

- ▶ Bayesian Approach:

- ▶ Bayesian encodes knowledge in the form of prior distributions.
- ▶ Conduct **prior predictive checks**:

$$P(\Theta) \rightarrow \Theta \rightarrow X$$

Checks if the data generated from the prior distribution aligns with realistic or expected values.

- ▶ Conduct **posterior predictive checks**:

$$P(\Theta | X) \rightarrow \Theta \rightarrow X^*$$

Evaluates how well the posterior distribution predicts new or observed data points.

- ▶ Rise of Computational Power:

- ▶ Advances in computational power have made simulations easier and more widely applicable.

## 03: Overparameterized Models and Regularization

- ▶ Adding parameters to a model allows for more flexibility and complexity:

### Model 1: Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- ▶ Explains the relationship between an outcome ( $Y$ ) and a predictor ( $X$ ).
- ▶ Limitation: Cannot account for time-varying effects or individual differences.

### Model 2: Incorporating More Parameters

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \text{time} + \beta_3 (X_i \cdot \text{time}) + \epsilon_i$$

- ▶ Adds parameters for time and its interaction with treatment, increasing flexibility.
- ▶ Still assumes uniform responses across individuals.

## 03: Overparameterized Models and Regularization (cont.)

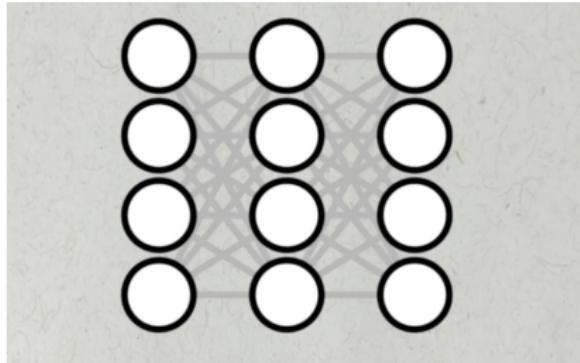
### Model 3: Mixed Effects Model (More Parameters, More Flexibility)

$$Y_i = \beta_0 + b_{0i} + \beta_1 X_i + b_{1i} X_i + \beta_2 \text{time} + \beta_3 (X_i \cdot \text{time}) + \epsilon_i$$

- ▶ Includes subject-specific random effects ( $b_{0i}, b_{1i}$ ), allowing for individual-level flexibility, increases model complexity.

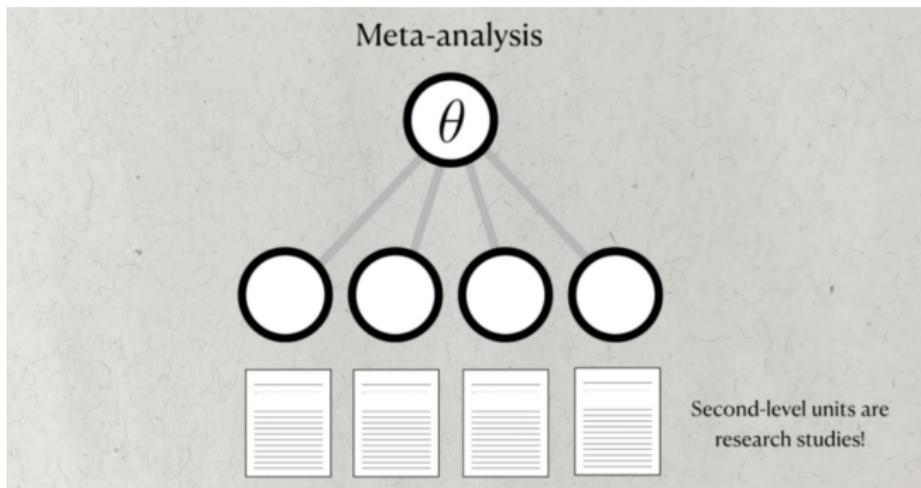
### Example: Neural Networks

- ▶ Overparameterized models take these models to the extreme, with hundreds or thousands of parameters.
- ▶ In neural networks, each edge in the network is associated with a parameter or weight. By making these networks very large, they can approximate a wide variety of functions (**Universal Approximation Theorem**).



## 04: Bayesian Multilevel Models

- ▶ Multilevel (or hierarchical) models assume a structure over parameters.
- ▶ These models allow for varying effects at different levels of hierarchy:
  - ▶ **First-level units:** Generate treatment effects.
  - ▶ **Second-level units:** Generate observed data and can vary depending on the research question.
- ▶ Example: Aggregating data across different groups in meta-analysis.



## 04: Bayesian Multilevel Models (cont.)

- ▶ The structure of the multilevel model can be expressed as:

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

- ▶ Definitions:

- ▶  $u_j$ : Random effect for group  $j$ .
- ▶  $\epsilon_{ij}$ : Error term for observation  $i$  in group  $j$ .

### Advantage

- ▶ Combines individual-level and group-level variability for improved estimates.
- ▶ The Bayesian framework offers more flexibility in the modeling approach, as the choice of prior distributions plays a major role in capturing uncertainties and incorporating domain knowledge.

## 05: Generic Computation Algorithms

### ► **Importance of Computers in Statistics:**

- ▶ Advances in computational power have been crucial for enabling and implementing complex statistical models.
- ▶ Algorithms like EM and Metropolis-Hastings make it possible to analyze data that would otherwise be computationally infeasible.

# The Expectation-Maximization (EM) Model

- ▶ **Purpose:** Solves estimation problems when data is incomplete or latent.
- ▶ **Process:** Uses available data to construct educated guesses for parameter values in a model ( $\hat{\theta} \rightarrow \theta$ ).
- ▶ **Example:** Mixture models with latent classes, where group labels are unknown. The EM algorithm estimates parameters even without direct class information.

# Metropolis-Hastings (M-H)

- ▶ **Purpose:** Allows sampling from complex distributions that are difficult to handle mathematically.
- ▶ **Bayesian Context:**
  - ▶ In Bayes' rule, the posterior distribution  $P(\theta | X)$  can be so complicated that we cannot derive a closed-form formula for it.
  - ▶ M-H generates data samples directly from this posterior, enabling us to:
    - ▶ Recover key quantities like means, quantiles, and credible intervals.
- ▶ **Importance:** Demonstrates the power of computation in Bayesian statistics, as it bypasses analytical challenges.

## 06: Adaptive Decision Analysis

- ▶ **Framework for Decision-Making:**
  - ▶ Interim decision points: Collect and analyze data before the experiment concludes.
- ▶ **Application:** Stopping clinical stem cell trials early:
  - ▶ If evidence suggests treatment is futile, the trial can be stopped early.
  - ▶ If treatment shows early promise, the trial may also be stopped to proceed with broader application.
  - ▶ A well-designed trial allows us to control the consequences of our decision, such as type I error.
- ▶ **Frequentist Considerations:**
  - ▶ Stopping early may hurt power and complicate the interpretation of the  $p$ -value.

## 07: Robust Inference

- ▶ Statisticians often make assumptions about their models. When assumptions are reasonable, results can generally be presumed trustworthy.
- ▶ **However, assumptions are not always right.**
- ▶ **Definition:** Robust statistics provide trustworthy analyses even when some assumptions are incorrect.

### Example: Outliers

- ▶ Outliers violate assumptions about the distribution of data (e.g., normality).
- ▶ Median-based estimators are less sensitive to outliers compared to mean-based estimators.

### Key Insight

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

# Propensity Score Matching

- ▶ **Definition:** A method to match people in a treatment group to people in a control group who are very similar to them.
- ▶ Propensity Score Matching requires two models:
  - ▶ **Model 1: Exposure-Outcome Model**

$$Y_i = \beta_0 + \beta_1 T_i + E_i$$

Estimates the effect of the treatment ( $T_i$ ) on the outcome ( $Y_i$ ).

- ▶ **Model 2: Propensity Score Model**

$$E_i = \tau_0 + \tau_1 C_i$$

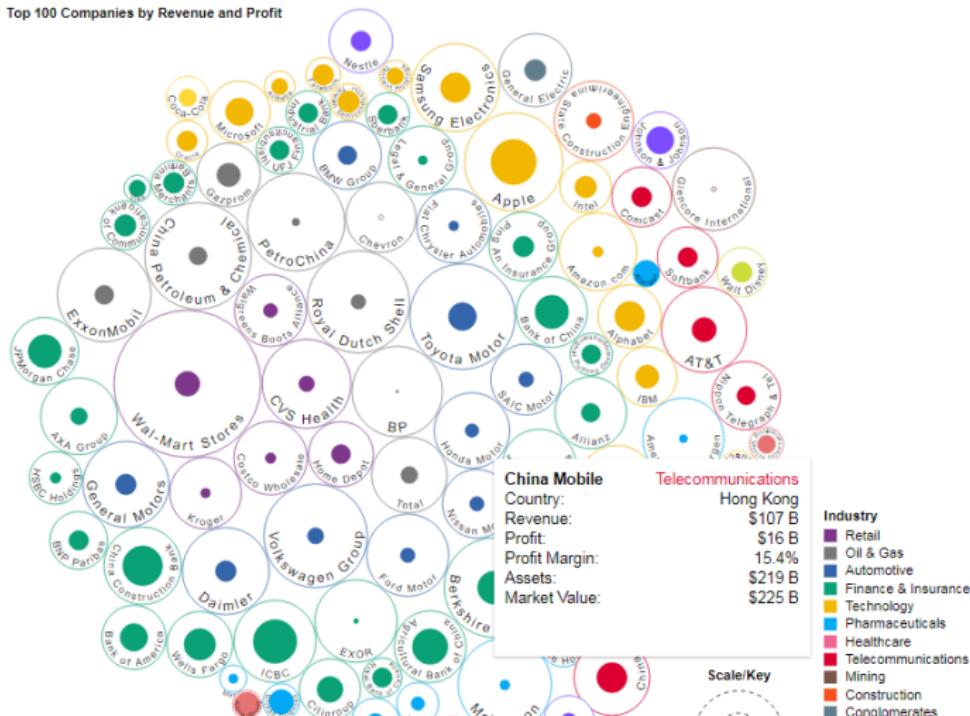
Matches individuals based on their covariates ( $C_i$ ).

- ▶ Both models must ideally be correctly specified; however:
  - ▶ Correct specification is rarely achieved in practice.
  - ▶ Robust versions of Propensity Score Matching allow one model to be misspecified.
- ▶ **Key Point:** The fewer assumptions we make, the better.  
Robust methods provide more reliable causal effect estimates.

## 08: Exploratory Data Analysis (EDA)

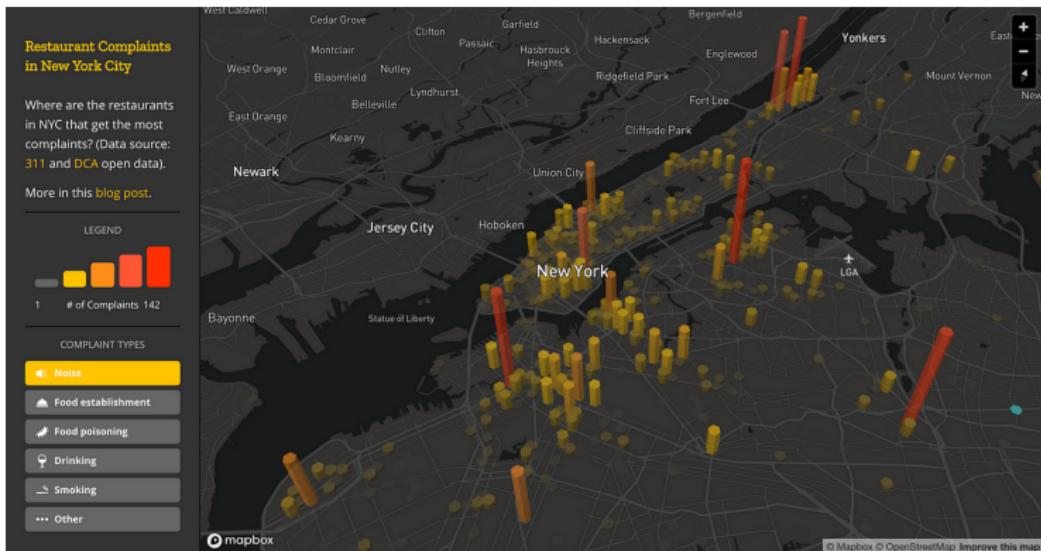
- ▶ Emphasizes visualization and insights over intense theory and computation.
- ▶ Useful in understanding the relation between data, fitted model, and predictions.

Top 100 Companies by Revenue and Profit



# Connection to NYC Open Data

- ▶ Apply statistical methods to NYC 311 and DCA open data.
- ▶ Example: Visualize and analyze restaurant complaints using robust inference and EDA.



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

# What does it mean for an idea to be important?

- ▶ Gelman emphasized avoiding citation counts of papers where concepts originated.
- ▶ Instead, he assessed principles that influenced the development of ideas in modern statistical practice.
- ▶ Example from the literature: The paper "The Importance of Being Clustered" (Anderluci et al., 2019) analyzed trends in statistics from 1970–2015 using clustering techniques to identify influential statistical ideas and themes.
- ▶ **My Remarks:**
  - ▶ Many of these ideas seem to have arisen from the analysis of very large datasets and the challenges inherent in combining sets of data.
  - ▶ The availability of computational power has also increased the tendency to model these statistical problems, pushing forward the development of modern methods.

# 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 in a separate interview

# 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(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>

## Images:

- ▶ bootstrap\_populations.png
- ▶ neural\_networks.png
- ▶ meta\_analysis.png
- ▶ propensity\_score\_matching.png
- ▶ bubble-chart.png