“What are the most important statistical ideas of the past 50 years?”(Andrew Gelman&Aki Vehtari) Analysis and sharing of findings after reading

Xuanming Dong

Author Introduction

Andrew Gelman is a professor of statistics and political science at Columbia University. He has received the Outstanding Statistical Application award three times from the American Statistical Association, the award for best article published in the American Political Science Review, the Mitchell and DeGroot prizes from the International Society of Bayesian Analysis, and the Council of Presidents of Statistical Societies award.

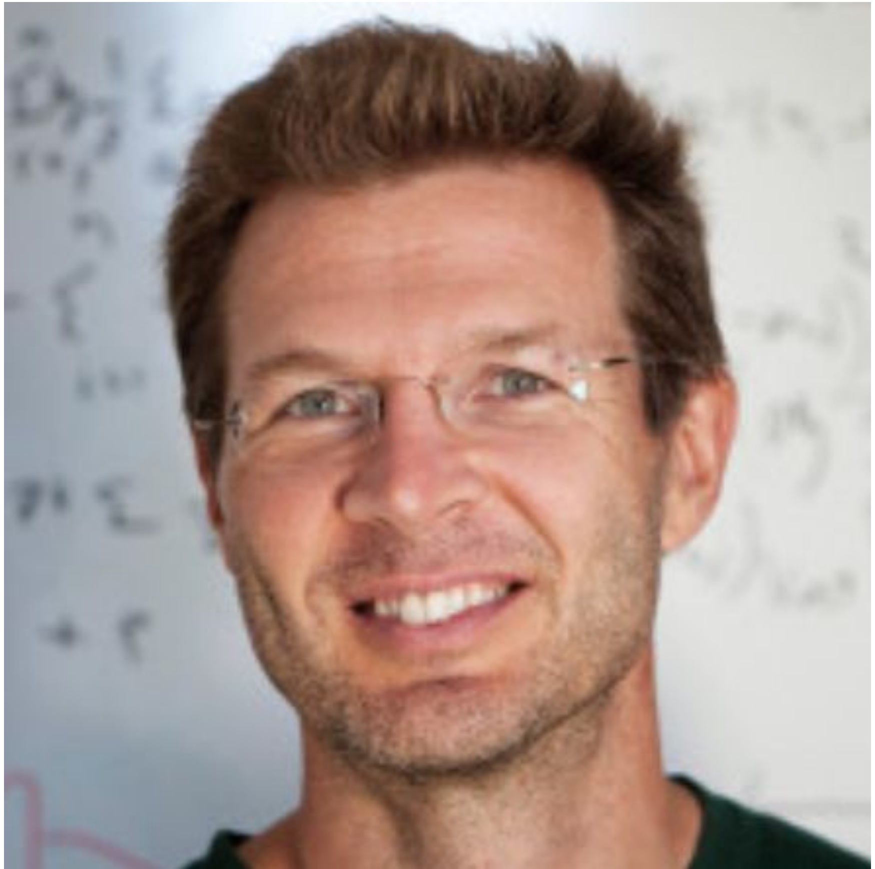
Author Introduction

Andrew has done research on a wide range of topics: such as, why it is rational to vote; why campaign polls are so variable when elections are so predictable; the effects of incumbency and redistricting; reversals of death sentences and etc,

In fact, we can learn from the above personal introduction that his contribution in statistics is impressive. At the same time, the scope of his statistical investigation is also very broad, ranging from some survey and statistical cases in life to some scientific research cases that we may not encounter in life.

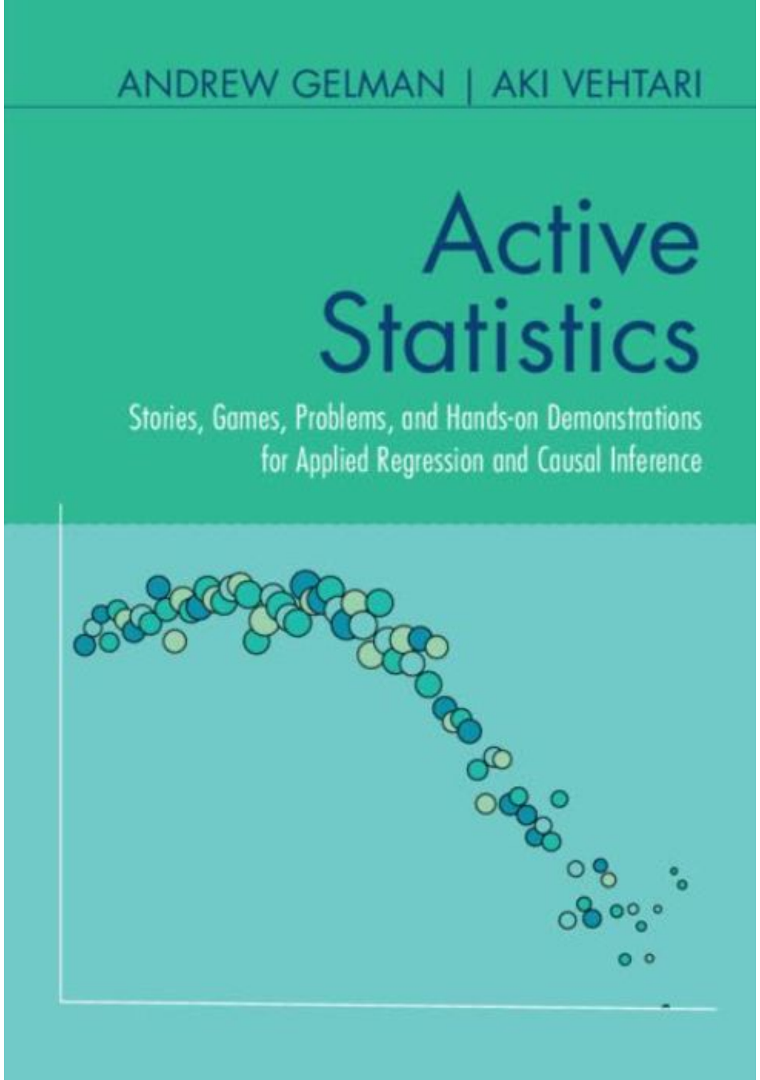
Author Introduction

Aki Vehtari, Professor of Computational Bayesian Modeling at Aalto University, Department of Computer Science.



Author Introduction

After checking some links, I found that Professor Aki Vehtari and Professor Andrew Gelman had a book called "Active Statistics" in 2024. This book mainly discusses statistical issues through 52 stories. It is very interesting that it is not just teaching through theoretical statements, but also using different cases as evidence of teaching. Moreover, after checking the classroom activities, I found that while teaching, it also combines the use of computers and data presentation analysis, which is very meaningful and can also allow students to understand what they have learned more intuitively.


ANDREW GELMAN | AKI VEHTARI
Active Statistics
Stories, Games, Problems, and Hands-on Demonstrations for Applied Regression and Causal Inference

Eight idears for the past 50 years

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

Counterfactual causal inference

Counterfactual causal inference refers to a method used to estimate the causal effect of an action, treatment, or intervention by comparing what actually happened (the factual outcome) to what would have happened in a hypothetical, alternative scenario (the counterfactual outcome).

For example, if there is a person who gets depressed if he or she does not have a pet, but does not get depressed when a pet is with her or him, then suppose we give this person a pet and the result is that this person will not get depressed. When we get this result (having a pet will not make that person depressed), how do we know from the beginning whether this person needs a pet's company?

Because we cannot conduct two experiments on the same experimental target, we cannot get a completely accurate answer, but we can use methods such as randomized experiments, matching, instrumental variables, etc. to approximate the counterfactual.

Counterfactual causal inference

Causal Effect= *Y1-Y0*

Y1 is the outcome under the treatment, Y0 is the outcome without the treatment.

so the Causal Effect can be whether the target have a pet.

In epidemiological and medical research, counterfactual or potential outcome models have gradually become the standard for causal inference, but this cannot provide 100% evidence of causality, but it is important to note that the reliability of the causal effect needs to be demonstrated in the research experiment.

Bootstrapping and simulation-based inference

The idea is to consider the estimate as an approximate sufficient statistic of the data and to consider the bootstrap distribution as an approximation to the sampling distribution of the data.

bootstrapping, The technique, which requires simple calculations, involves drawing repeated samples (with replacement) from the empirical–or the actual–data distribution and then building a distribution for a statistic by calculating a value of the statistic for each sample.

Overparameterized models and regularization

the idea of fitting a model with a large number of parameters—sometimes more parameters than data points—using some regularization procedure to get stable estimates and good predictions. The idea is to get the flexibility of a nonparametric or highly parameterized approach, while avoiding the overfitting problem.

For example, a person needs to buy a car, but there are many parameters that will affect the price of the car, such as the brand, usage time, driving distance, etc. A person may need to consider dozens or even hundreds of different factors before buying a car. Many different parameters may lead to overfitting of the price of the car, so we need to regularize some data that the customer does not need, such as the tire model of the car.

Bayesian multilevel models

Bayesian multilevel models are a type of statistical model that allows for the estimation of parameters at multiple levels of a hierarchy using Bayesian methods. These models are particularly useful for analyzing data that has a grouped or nested structure, The "multilevel" aspect refers to the idea that the data can be organized into different levels, and each level can have its own set of parameters.

For example, if the government decides to investigate the eating habits of teenagers in school, it can summarize and conclude based on multiple factors such as individual teenagers, different schools, different age groups, and different states. This will better distinguish the differences in the research subjects at different levels.

Generic computation algorithms

algorithms provide general solutions for tasks like sorting, searching, optimization, and numerical computations, and they are highly adaptable to a wide range of specific applications across multiple fields.Generic computation algorithms form the backbone of solving various problems in computer science, mathematics, and engineering. These algorithms provide general solutions for tasks like sorting, searching, optimization, and numerical computations, and they are highly adaptable to a wide range of specific applications across multiple fields.

For example, the GPS algorithm in map positioning. When we want to go from starting point A to destination B, this type of algorithm can help us set up different nodes, and by updating maps, weather, environment and other factors in real time, design the shortest distance from A to B for us.

Adaptive decision analysis

Adaptive Decision Analysis refers to a systematic approach to decision-making that adjusts or "adapts" based on evolving information, changes in the environment, or feedback from previous decisions. This approach is useful in complex, uncertain, or dynamic situations where the optimal decision may not be clear from the outset and may change over time as new information becomes available.

The key concepts of this decision-making include analysis and feedback of uncertainty, learning and analysis, etc. A good example is the development and extension of artificial intelligence. For example, if a new electric car wants to have an automatic driving system, then Adaptive decision analysis will help the artificial intelligence deal with what may happen to the car at any time, such as road traffic conditions, route planning, etc. In theory, the driver does not even need to make decisions for his own driving. The adaptability, analysis and decision-making of artificial intelligence can help the car work safely and orderly.

Robust inference

The idea of robustness is central to modern statistics, and it’s all about the idea that we can use models even when they have assumptions that are not true. An important part of statistical theory is to develop models that work well, under realistic violations of these assumptions.

For example, if we want to predict the return rate of certain stocks in the stock market, from a traditional perspective, the model may have problems due to possible financial crises, epidemics, etc., which cannot be fully explained by analyzing indicators such as interest rates and inflation rates. Then we can reorganize our outlook and return rate of stocks by correcting the error values in different robust inferences.

Exploratory data analysis

Exploratory Data Analysis (EDA) is an approach to analyzing and summarizing datasets by using statistical graphics and visualization techniques, along with simple descriptive statistics. The goal of EDA is to understand the underlying structure of the data, detect anomalies, test hypotheses, and check assumptions before applying more complex modeling techniques.

We can use R to observe the data more intuitively, and from the joint display of pictures and data, we can also more clearly understand the authenticity and effectiveness of the presented data. (such as histograms, box plots, etc.)

Have in common and differ

The eight different ideas have changed the workflow in some sense, and with the help of computer technology, calculations are better, and there are also powerful data and different types of models for experiments and research. At the same time, these ideas are related to each other. For example, the hierarchical and data-based Bayesian multi-layer model can be displayed in the form of EDA, which is more intuitive. At the same time, the predictability and decision-making of the uncertainty of the results are constantly derived and developed.

Even though the fields they involve are different in actual jobs, there is no ranking difference between different ideas. I hope that what we explore is not the ranking of different ideas or to emphasize the importance of different ideas, but that we can explore the products that may be derived from one of the ideas.

For the future

Maybe in the future, as the author said, "some of the most important statistical research of the next fifty years will lie at the interface of high-dimensional and nonparametric modeling and computation on one hand, and causal inference and decision making on the other." At the same time, changes in the times and technology also allow us to use more different ways to analyze and observe more medium-parameter data, so as to obtain more effective and meaningful views. In different modules, data and ideas, we can find their connections, which are inevitable. Through the connections in different ideas, we may be able to use more advanced means to calculate, summarize, analyze, more parameters and meaningful goals.

Reference page

Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University. (n.d.). http://www.stat.columbia.edu/~gelman/

Gelman A. & Vehtari A. (2021). What are the most important statistical ideas of the past 50 years? Journal of the American Statistical Association, 116(536): 2087–2097.

Active statistics. (n.d.). https://avehtari.github.io/ActiveStatistics/