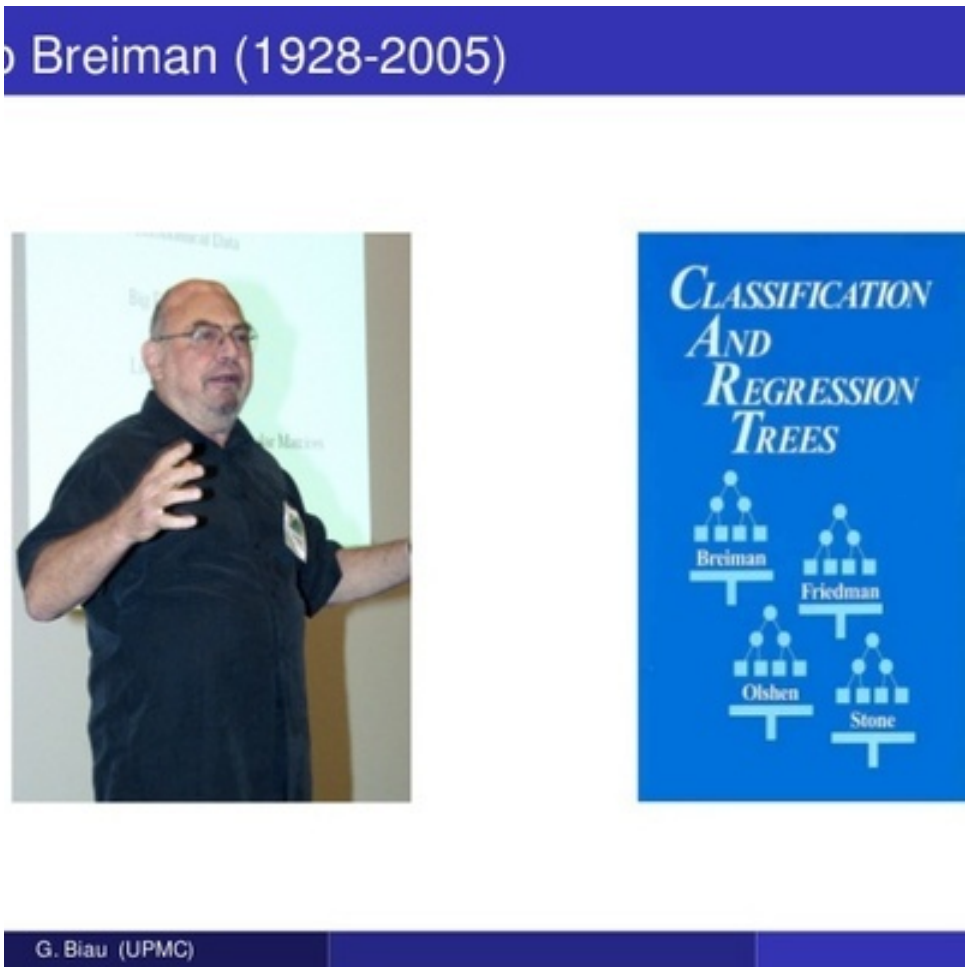

Breiman's World
Where's the randomness

Stat Student

2024-04-27

Contents

Dedication

To my instructors who dedicated their life to learning and teaching the art of data analysis: TSP, LB, DAF.

Chapter 1

Introduction

This book is a compilation of lecture notes in applied statistics from a course taught by Leo Breiman at the UC, Berkeley statistics department in 1991, 10 years before the publication of the seminal article - **Statistical Modeling: The Two Cultures** [1]. Peppered throughout the notes are indications of Pofessor Breiman's struggle to find among the available statistical tools helpful resources for some of the practical data analysis problems he was up against. **Statistical Modeling: The Two Cultures** [1] is a good read to have before going through these notes.

Another interesting read to have before assaulting these notes is **Nail Finders, Edifices and Oz** [2] in which Professor Breiman expresses great dissatisfaction with the way statistics was being taught at the time, leaving statsisticians ill-equipped to solve real-world problems.

Finally, another good read is **50 Years of Data Science** [3] (the first few sections), in which Breiman's urging academic statistics to expand its boundaries beyond the classical domain of theoretical statistics is recognized along with similar efforts coming from contemporaries John Chambers [4], Jeff Wu (slides here), and Bill Cleveland [5] as well as from their predecessor John Tukey [6][1] , some 50 years before them.

1.1 Where's the randomness?

Against this background of the need for change, when we enter Leo Breiman's classroom we enter another world, one deeply steeped in the theoretical underpinnings of applied statistics. Although all of the necessary theory was included in the course material, we spent very little time marvelling at the beauty of

[1] Analysts at GRAIL should recognize John Tukey as the inventor of the most utilized data display tool in the company

the mathematical results underlying some statistical theories (for example ...) and instead were encouraged and challenged to struggle with difficult questions: what happens if this assumption is not true? or that assumption? I remember one particular lecture watching Professor Breiman incessantly pacing up and down the front of the class, asking "**where is the randomness? where's the randomness coming from**?" I also remember not putting much of an effort in providing an answer [2].

Some 30 years later, faced with the problem of trying to answer questions about a system generating data which comes in the form of batches of points having a yet to be fully understood correlation structure and subject to sporadic surges of variation of unknown origin [3]. I now fully appreciate the relevance of the question - **where's the randomness coming from**. I don't think there is much any exaggeration in claiming that if we can identify the relevant sources of variability, and how to capture them in an analysis dataset, we have basically completely solved the data analytic problem: we know exactly what data to collect and how to summarize them in order to answer the question of interest.

1.2 What next?

We will go through professor Breiman's notes in much the same order as we did in class, and try to connect the principles and questions raised to problems we face, thinking about the data and sources of variability - where is the randomness coming from. If we know this, we know everything we need to know. My hope is that having gone through these notes, we will be convinced of the truth of this idiom, and this in turn will provide the drive necessary to get to a full and accurate knowledge of the randomness present in the task at hand. Equipped with this knowledge we will have full confidence that the results are interpretable as intended and surprises becoming the exception rather than the rule.

Professor Breiman's class notes provide the statistical theory background necessary to move forward with most data analysis projects. This may be a minimalists' point of view in terms of required statistical theory background; we will see as we work through some problems.

1.3 Nail Finders

It would be good to start with a number of concrete questions which we can use as exemplar test cases to keep in the back of our minds as we go through the notes. We must avoid the usual textbook format of having neat examples associated with each introduced technique. This is exactly the format which leads to the belief that there is a ready-made canned solution for every problem. In this

[2] I was not a very good student, to say the least and was told number of times that I should exchange my student visa for a tourist visa ...

[3] This variations is what we refer to as batch effects, or run to run variability and encompasses all unexplained variability shared by a set of samples

modus operandi, little to no effort is spent trying to get to a deep understanding of the data and the problem at hand as one only tries to characterize the problem to the extent necessary to find a suitably well matched canned solution. Such trivial examples do exist and for the most part these do not require much statistical expertise; we can leave those to AI, or Chad Gpt.

1.4 The edifice complex

Another trap now presents itself - inventing complexity when it isn't there. While many of the problems which present themselves in the course of our work will lend themselves to analysis using traditional and more or less standard linear modeling statistical techniques, many, if not all, will have unique features - randomness in the data, application or question of interest, and inferential frame. Identifying the unique features present of the task at hand is the primary role of the statistical analyst. If there are no new features to a new problem, leave it to AI or change the link to the input in last weeks' code, push go, and turn your attention to the next problem which requires careful consideration and may call for some innovation and ingenuity.

1.5 Oz

Chapter 2

Preliminaries

Before we get into Brieman's world we should mention some important elements of data analysis which are not part of the class notes: Exploratory Data Analysis and Data Simulation.

2.1 Exploratory Data Analysis

> Measure 7 times, cut once.
>
> — Russian saying

> Verify all assumptions. Look before, and after, you transform. When you transform, verify!
>
> — Uncommon Sense

Exploratory data analysis refers to a set of tools which enable us to navigate the Uncommon Sense principles. A concise history is [7]:

> Exploratory data analysis is a set of techniques that have been principally developed by Tukey, John Wilder since 1970. The philosophy behind this approach is to examine the data before applying a specific probability model. According to Tukey, J.W., exploratory data analysis is similar to detective work. In exploratory data analysis, these clues can be numerical and (very often) graphical. Indeed, Tukey introduced several new semigraphical data representation tools to help with exploratory data analysis, including the "box and whisker plot" (also known as the box plot) in 1972, and the stem and leaf diagram in 1977. This diagram is similar to the histogram, which dates from the eighteenth century.

References are to [8] and [9].

Exploratory data analysis is a topic that is not treated in Professor Breiman's notes, but cannot go un-mentionned when discussing and taking instructions on data analysis and modeling. Being aware of the salient features of the data is necessary before any statistical modeling and analysis can take place. Skip this step and you will get into trouble, Guaranteed!

EDA is greatly facilitated by readily available software as well as the modern hardware technologies which can quickly render beautiful plots. Trevor Hastie, in a talk which is further explored below (Trevor Hastie SLBD Talk (Bristol 2018) and accompanying SLBD Slides), notes that with the advent of big data, data visualization has become challenging again and some may be tempted skip that essential step of the data analysis process. Professor Hastie remarks that the problems one runs into when analyzing data without familiarising oneself with the salient features of the data may be harder to detect but the have not disappeared.

We will not fill this chapter with descriptions of EDA summaries and plots. The methods are varied and many (See [9–11]), and their utility intimately dependent on the context. EDA methods will instead be introduced as they come up in the course of the analyses described below.

2.2 Simulated Data Experimentation

> when you simulate fake data, you kinda have to have some sense of what's going on. You're starting from a position of understanding.
>
> — Andrew Gelman

Simulations to evaluate assumptions made about the data structure, and for verifying statistical procedures and properties of model fits is another essential part of the data analysis process which didn't make it into Professor Breiman's notes. Andrew Gelman is nuts about fake-data simulation and can't stop blogging about it:

- Yes, I really really really like fake-data simulation, and I can't stop talking about it.
- Simulated-data experimentation: Why does it work so well?
- Why I like hypothesis testing (it's another way to say "fake-data simulation")
- Simulation to understand measurement error in regression - to appear
- the power of fake data simulations - this blog, with code, is a replication of what Andrew Gelman has already posted - Multilevel data collection and analysis for weight training (with R code).

2.2.1 To do - summarize the main points of Gelman's work

One take home will be - it's not easy to achieve an appropriate level of similarity between the fake data and the actual data to be analyzed. In software engineering tools, to each fixed input a specific output is expected and when there are no abends or errors of any kind and the expected output is obtained, the task is basically complete. In statistical analysis and data simulations, once the code works the task has only begun. Once the simulation code "works", we need to simulate enough fake data to see if it really works - does it capture all of the salient features of the real data.

In addition to his work on data simulations, Andrew Gelman has also published on the topic of `multilevel data` which may be analogous to the data structure of plated samples coming off the pipeline. See Gilman and Hill (2006) [11], which also has a chapter on simulated data [12] and software to boot (See Data Analysis Using Regression and Multilevel/Hierarchical Model). Other works of interest are Gelman, Hill and Vehtari (2020) [13] and not yet published "Advanced Regression and Multilevel Models".

Chapter 3

What is statistics?

3.1 The data modeling approach

> to pull a rabbit from a hat, a rabbit must first be placed in the hat
>
> — David A. Freedman

3.1.1 Probability

George Box famously said [14]:

> all models are wrong, but some are useful

A far better quotation from the same paper by Box is,

> Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers around[1]
>
> — Phil Stark (2022) [15]

3.1.2 Statistics

3.2 The predictive modeling approach

> Without data, you're just another person with an opinion
>
> — W. Edwards Deming

What Professor Breiman referred to as `the algorithmic approach` was later called 'the predictive modeling approach by David Donoho [3]. We will adopt that terminology as well.

[1] slight modification of quote

3.3 The predictive data modeling approach

Some modern data problems which involve the analysis of large datasets can only be solved by a combination of both data and predictive modeling approaches. Omics - genomics, transcriptomics, proteomics, metabolomics and other omics - refer to high-throughput experimental technologies which produce measurements related to the molecular content of cells. Genomic platforms, for example, record measurements indicating the abundance of RNA transcripts present in the cells of processed samples. Different platforms use different technologies to obtain the measurements on transcript abundance, and all platforms require sophisticated pipelines to process the platform's hardware readout and produce the transcript abundance measurements required for downstream analysis. These in turn require further processing in order to provide data in a form that can be used to address the biological questions of interest, which genes are differentially expressed between two groups of samples for example.

First we discuss data analyses which are carried out at the so-called down-stream end of the analysis pipeline, following earlier analyses, sometimes called pre-processing, which transform the machine read-outs into analysis units - vectors of gene expression and protein concentrations, for example.

Hastie and colleagues coined the term `Statistical Learning` to refer to[2]:

> a branch of applied statistics that emerged in response to machine learning, emphasizing statistical models and assessment of uncertainty.

The primary contrast with `Machine Learning`, which is a field focused on constructing algorithms that can learn from data, is the incorporation of assessment of uncertainty in the evaluation of a predictor's performance[3]. The statistical learning approach is prescribed and illustrated in [16–19].

3.3.1 Example: microarray data

[2] this material taken from Trevor Hastie SLBD Talk (Bristol 2018) and accompanying SLBD Slides

[3] I can't believe that machine learners have no interest in uncertainty. In this story the contrast must be between an assessment of uncertainty based entirely on empirical results, as used by the machine learning community,a and an assessment of uncertainty which incorporates some theoretical results in addition to empirical findings. I think that this distinction is rather slim. Another distinction that could be made is the we can assume that the statistical learner, by virtue of their desire to incorporate theoretical results in the assessment of uncertainty, will have a preference for simpler models, which may translate into a stronger sense of generalizability and extendability.

Chapter 4

Linear Models

> We were together learning how to use the analysis of variance, and perhaps it is worth while stating an impression that I have formed-that the analysis of variance, which may perhaps be called a statistical method, because that term is a very ambiguous one — is not a mathematical theorem, but rather a convenient method of arranging the arithmetic. Just as in arithmetical textbooks — if we can recall their contents — we were given rules for arranging how to find the greatest common measure, and how to work out a sum in practice, and were drilled in the arrangement and order in which we were to put the figures down, so with the analysis of variance; **its one claim to attention lies in its convenience.**
>
> The Future of Data Analysis
>
> — John Tukey

4.1 Regression

4.2 Analysis of Variance

Chapter 5

Design of Experiments

Chapter 6

Generalized Linear Models

6.1 Binary Data and Logictic Models

6.2 Count Data and Log-Linear Models

Chapter 7

Predictive Analysis

7.1 Traditional Classification Analysis

7.2 Modern Predictive Modeling

7.2.1 CART - The Birth of a Methodology

- Faced with a new question and data of an unsual format, professor Breiman went on to develop a new methodology which would eventually give rise to ...

Chapter 8

Appendix

8.1 Breiman versus Cox

In Statistical Modeling: The Two Cultures [1] Professor Leo Breiman bemoans the fact that a vast majority of statisticians have pigeon-holed themselves to such an extent that the field is becoming irrelevant with the emergence of new data and problems which do not conform to the requirements of classical statistical methodologies. When Professor Breiman returned to academia in 1980 after 13 years of statistical consulting, he was deeply troubled by the prevalent modus operandi in the field of statistics: *every article in the Annals of Statistics, the flagship journal of theoretical statistics, started with* **Assume that the data are generated by the following model**, followed by mathematics exploring inference, hypothesis testing and asymptotics. Professor Breiman goes on to describe some of the deficiencies in the applications of statistics which this approach has led to before going on to describe an alternative to the classical statistical data modeling approach, `algorithmic modeling`[1].

There are two parts to Professor Breiman's criticism. One part of the criticism is that some modern day problems are just not amenable to the classical statistical approaches and require a shift in the conventional statistical data analysis paradigm. The other part of the criticism is that even when data modeling is an appropriate way to solve a problem, the traditional approach has led to questionable analyses, results, and conclusions due to uncritical use of standard methods and lack of effort to customize analysis to the specific needs of each problem.

One of the statisticians invited to comment on Professor Breiman's criticism of the current state of affairs in the field of statistics was Sir David Cox, known for, among many other influential works, the Cox proportional hazards model.

[1] Donoho [3] pefers the term `predictive modeling`

We will focus on his comments as his disagreement with the notions put forth professor Breiman are the strongest among the commentators.

- "Professor Breiman takes data as his starting point. I would prefer to start with an issue, a question or a scientific hypothesis"
  - Nobody would seriously argue that starting with anything other than the question is a good idea, but the tendency of practitioners to assume convenient data models, and use the canned, turn-key analytical procedures which come with these models, suggests that these practitioners are quite content with starting with the solution, which one might argue is an even bigger assault to common sense than starting with the data.

- "Professor Breiman emphasizes prediction as the objective, success at prediction being the criterion of success, as contrasted with issues of interpretation or understanding".
  - This is another moot point. When we can confidently proceed with a defendable data model, the assessment of the utility of predictions made with the model can only enhance interpretation.

- "Often the prediction is under quite different conditions from the data. ie. it may be desired to predict the consequences of something only indirectly addressed by the data available for analysis."
  - Making predictions to a space beyond the space spanned by the data used to estimate the model would be hazardous, no matter what modeling approach is used.

- After describing an idyllic data analysis package and explaining why the details of such are not published, Sir David goes on to say "By contrast, Professor Breiman equates mainstream applied statistics to a relatively mechanical process involving somehow or other choosing a model, often a default model of standard form, and applying standard methods of analysis and goodness-of-fit procedures. ...
  It is true that many of the analyses done by non-statisticians or by statisticians under severe time constraints are more or less like those Professor Breiman describes."
  - Professor Cox admits that this automated or mechanical process is likely to be favored when working under severe time constraints;
    - What Sir David Cox doesn't realize is that this is part of the job description for a statistician working in industry - must be prepared to work under severe time constraints. I am a witness to the fact that the relatively mechanical process described by Professor Breiman **is** mainstream applied statistics in my industry[2].

[2] I am not prepared to put the blame entirely on the lack of preparation statistics graduates might have as time on the job does not seem to translate to different behaviours: more careful examination of the questions and the development of methods suited to the particulars of the problems. Some blame has to be attributed to the system in place. A system in which there

- Professor Cox goes on to forgive the laps in rigor in the application of statistical techniques to analyse data with the observation that "one suspects that quite often the limitations of conclusions lie more in weakness of data quality and study design than in ineffective analysis." How would a practitioner who uncritically uses the common modern day approach ever become aware of the weakness of the data or the design? The answer is likely to be never, as careful consideration is rarely given to the suitability of the model or the quality of the data.

8.2 Freedman on John Snow, Shoe Leather and Lines of Evidence

Notes

- [20]:
  - Snow's work on cholera is presented as a success story for scientific reasoning based on nonexperimental data. Failure stories are also discussed, and comparisons may provide some insight. In particular, this paper suggests that statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings.
- [21]:
  - Proportional-hazards models are frequently used to analyze data from randomized controlled trials. This is a mistake. Randomization does not justify the models, which are rarely informative. Simpler analytic methods should be used first.

Also see [27].

1. Breiman, L. (2001). Statistical modeling: The two cultures. Statistical Science *16*, 199–215. Available at: http://www.jstor.org/stable/2676681.
2. Breiman, L. (1984). Nail finders, edifices, and oz (Department of Statistics, University of California) Available at: https://books.google.com/books?id=y5-HHAAACAAJ.
3. Donoho, D. (2017). 50 years of data science. Journal of Computational and Graphical Statistics *26*, 745–766. Available at: https://doi.org/10.1080/10618600.2017.1384734.
4. Chambers, J.M. (1993). Greater or lesser statistics: A choice for future research. Statistics and Computing *3*, 182–184. Available at: https://api.semanticscholar.org/CorpusID:122235676.
5. Cleveland, W. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. International Statistical Review / Revue Internationale de Statistique *69*.

are no checks in place to ensure the lasting validity of results but rather only checks on the timeliness of the delivery of results, provides little incentive for data analysis professionals to incorporate more careful scrutiny of assumptions or innovative solutions in their practice

6. Tukey, J.W. (1962). The future of data analysis. The Annals of Mathematical Statistics *33*, 1–67. Available at: http://www.jstor.org/stable/2237638.

7. Exploratory data analysis (2008–). In The concise encyclopedia of statistics (New York, NY: Springer New York), pp. 192–194. Available at: https://doi.org/10.1007/978-0-387-32833-1_136.

8. Tukey, J.W. (1972). Some graphic and semigraphic displays. Statistical papers in honor of George W. Snedecor *5*, 293–316.

9. Tukey, J.W. (1977). Exploratory data analysis (Springer).

10. Cleveland, W.S. (1993). Visualizing data (Murray Hill, NJ: AT & T Bell Laboratories) Available at: https://dl.acm.org/doi/abs/10.5555/529269.

11. Gelman, A., and Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models (Cambridge: Cambridge University Press) Available at: https://www.cambridge.org/core/product/32A29531C7FD730C3A68951A17C9D983.

12. Gelman, A., and Hill, J. eds. (2006). Simulation for checking statistical procedures and model fits. In Data analysis using regression and multilevel/hierarchical models Analytical methods for social research. (Cambridge: Cambridge University Press), pp. 155–166. Available at: https://www.cambridge.org/core/product/9546B05E0BE3DD5F20F63995F4139198.

13. Gelman, A., Hill, J., and Vehtari, A. (2020). Regression and other stories (Cambridge: Cambridge University Press) Available at: https://www.cambridge.org/core/product/DD20DD6C9057118581076E54E40C372C.

14. Box, G.E.P. (1976). Science and statistics. Journal of the American Statistical Association *71*, 791–799. Available at: http://www.jstor.org/stable/2286841.

15. Stark, P.B. (2022). Pay no attention to the model behind the curtain. Pure and Applied Geophysics *179*, 4121–4145. Available at: https://doi.org/10.1007/s00024-022-03137-2.

16. Hastie, T., Tibshirani, R., and Friedman, J.H. (2009). The elements of statistical learning: Data mining, inference, and prediction (Springer) Available at: https://books.google.com/books?id=eBSgoAEACAAJ.

17. Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical learning with sparsity: The lasso and generalizations.

18. Taylor, J., and Tibshirani, R.J. (2015). Statistical learning and selective inference. Proceedings of the National Academy of Sciences *112*, 7629–7634. Available at: https://doi.org/10.1073/pnas.1507583112.

19. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2023). An introduction to statistical learning: With applications in r G. James, D. Witten, T. Hastie, and R. Tibshirani, eds. (New York, NY: Springer New York) Available at: https://www.statlearning.com/.

20. Freedman, D.A. (1991). Statistical models and shoe leather. Sociological Methodology *21*, 291–313. Available at: http://www.jstor.org/stable/270939.

21. Freedman, D. (2008). Randomization does not justify logistic regression. Statistical Science *23*.

22. Freedman, D. (2009). Diagnostics cannot have much power against general alternatives. International Journal of Forecasting *25*, 833–839.

23. Freedman, D. (1999). From association to causation: Some remarks on the history of statistics. Statistical Science *14*, 243–258. Available at: https://doi.org/10.1214/ss/1009212409.

24. Freedman, D., Petitti, D., and Robins, J. (2004). On the efficacy of screening for breast cancer. International journal of epidemiology *33*, 43–55.

25. Freedman, D.A. (2009). On types of scientific inquiry: The role of qualitative reasoning. In Statistical models and causal inference: A dialogue with the social sciences, D. A. Freedman, D. Collier, J. S. Sekhon, and P. B. Stark, eds. (Cambridge: Cambridge University Press), pp. 337–356. Available at: https://www.cambridge.org/core/product/19AD82B406B3F0B1EAF12B4A6AC33E23.

26. Freedman, D. (2009). Statistical models: Theory and practice.

27. Freedman, D.A. (2008). Survival analysis. The American Statistician *62*, 110–119. Available at: https://doi.org/10.1198/000313008X298439 .