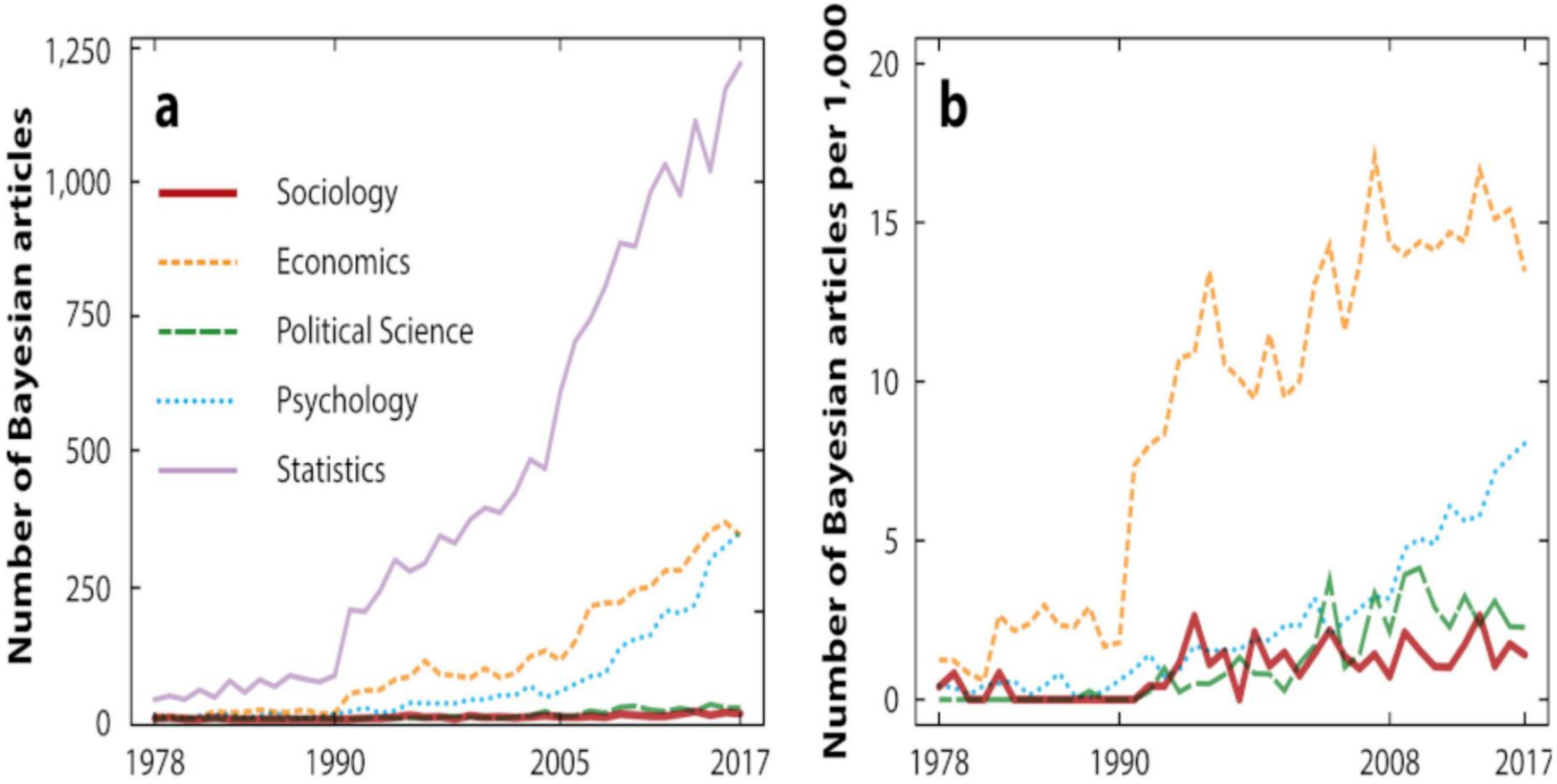
GR5065: Introduction

Ben Goodrich

January 24, 2022

Bayesian Articles (from Lynch and Bartlett 2019)



Undergraduate Bayesian Courses

- One of the coauthors of the *Bayes Rules* textbook also coauthored a paper on the prevalence of Bayesian courses at 102 U.S. universities and 50 liberal arts colleges (see also this blog post which discusses the paper)
- Of these, 45 universities and 6 liberal arts colleges have Bayesian courses, almost all of which are conducted by the statistics or math departments
- Only 4 require a Bayesian course for the major
- The 51 Bayesian courses tend to have about SIX prerequisite courses in calculus, linear algebra, computing, statistics, and probability
- Almost all students in GR5065 will not have taken an undgraduate class in Bayesian statistics

Quotes from page 3 of *Statistical Rethinking*

- “statistics is neither mathematics nor a science, but rather a branch of engineering”
- “Advanced courses in statistics do emphasize engineering, but most scientists never get that far. Teaching statistics this way is like teaching engineering backwards, starting with bridge building and ending with basic physics.”
- “Serious trouble begins when scholars move on to conducting innovative research, pushing the boundaries of their specialties. It’s as if we got our hydraulic engineers by promoting plumbers.”
- “Why aren’t the tests enough for innovative research? The classical procedures of introductory statistics tend to be inflexible and fragile. By inflexible, I mean that they have very limited ways to adapt to unique research contexts. By fragile, I mean that they fail in unpredictable ways when applied to new contexts.”

What Is GR5065 About?

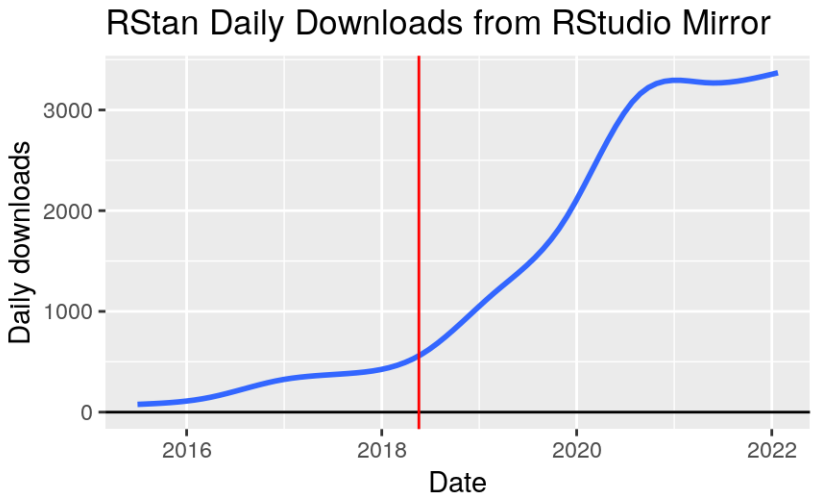
- The analogue of “basic physics” for us is probability: a shared language for communicating about uncertain (sometimes future) propositions
- QMSS does not offer / encourage a probability course such as GU4203
- GR5065 is essentially a combination of:

1. A full semester of probability at the master’s level
2. A substantial part of theory & methodology that was not covered in the theory & methodology course, GR5010, QMSS students are required to take
3. Learning new R packages that utilize Stan to put (1) and (2) into practice
4. Unlearning much of what you thought you learned in other classes

Obligatory Disclosure

- Ben is an employee of Columbia University, which has received several research grants to develop Stan
- Ben is also a manager of GG Statistics LLC, which uses Stan for business
- According to Columbia University policy, any such employee who has any equity stake in, a title (such as officer or director) with, or is expected to earn at least $\$5,000.00$ per year from a private company is required to disclose these facts in presentations

0:00 / 0:19



Season 3, Episode 9 of Billions

R and Stan

- This course uses the statistical software R(Studio)
- If you have neither used R nor another programming language (e.g. Python, Java, C++) this course will be VERY difficult because we do not teach the basics of R since QMSS students have already used R for at least one semester
- Homeworks have to be done using RMarkdown (we will explain this)
- Stan is another programming language, which we will not learn directly but can be accessed from a variety of other interfaces, besides R which you might prefer to use for Bayesian inference after the course is finished

The Simplest Impossible Problem

Collatz Conjecture: R function below should return TRUE for all positive integers:

```
is_even <- function(x) (x %% 2) == 0
Collatz <- function(x) { # x must be a single (representable) positive integer
  stopifnot(length(x) == 1, is.numeric(x), is.finite(x), x > 0, x == floor(x))
  while (x != 1) {
    if (is_even(x)) {
      x <- x / 2
    } else {
      x <- 3 * x + 1
      if (is.infinite(x)) return(NA) # overflow so conjecture is undetermined
    }
  }
  return(TRUE) # because x is currently 1
}
```

If there were ANY positive integer where this sequence of x values diverges to ∞ or enters a cycle not involving 1, then the Collatz Conjecture would be false.

Aristotelian (Classical Propositional) Logic

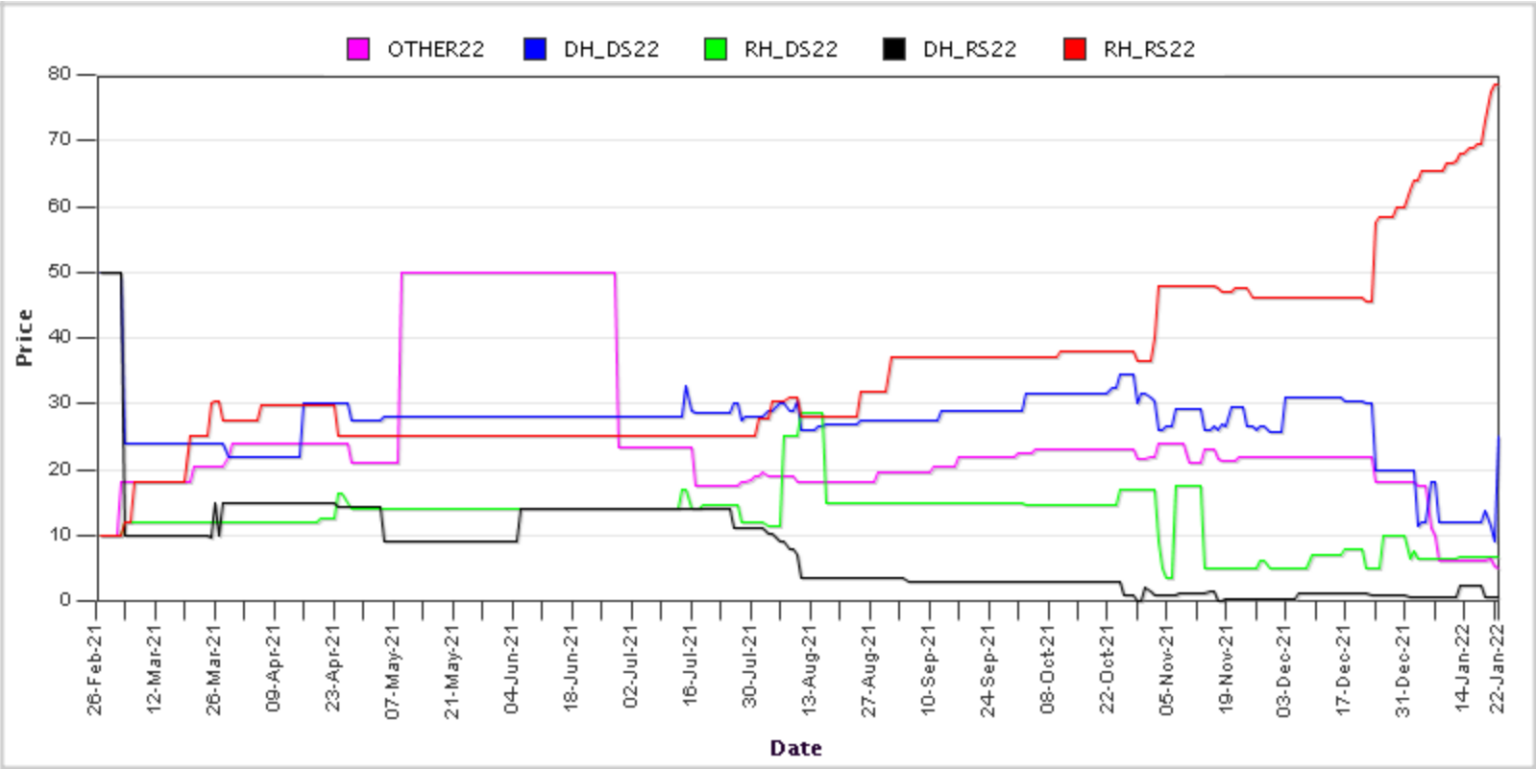
1. All men are mortal
2. Socrates is a man
3. Ergo, Socrates is mortal

- There are 0 interesting applications of deductive logic in the social sciences
- The closest is perhaps democratic peace “theory”:
  1. No two democratic countries will fight a war against each other
  2. Australia and Serbia are democracies
  3. Ergo, Australia and Serbia will not fight a war against each other
- Whether (1) is true depends on how you operationalize “country”, “democracy”, and “war” so the literature has de-evolved into debates over how those terms should be operationalized
- A better (probabilistic) question is: Why are democracies much less likely to fight a war against each other compared to wars involving an autocracy?

Probability as an Extension of Aristotelian Logic

- Various people have characterized probability as a weaker form of logic where we are not necessarily certain whether propositions are true or false
- In R and many other programming languages, `TRUE` maps to 1 and `FALSE` maps to 0 for the purpose of doing calculations. Probabilities are all the real numbers between 0 and 1, and we can proceed with (probabilistic) inference.
- An example:
  1. The party of the U.S. president TENDS to lose Congressional seats in midterm elections
  2. Joe Biden is a Democrat (which in this example we are certain of)
  3. Democrats will PROBABLY lose Congressional seats in the 2022 election
- Neither (1) nor (3) is guaranteed, although (3) seems plausible — to some degree — to the extent you accept both (1) and (2)
- Bayesianism is a school of thought that uses probability to describe the degree of belief (with quantified uncertainty) that a proposition is true

Iowa Electronic Market for 2022 Congress



Four or Five Sources of Uncertainty

1. Uncertainty about parameters in models
2. Uncertainty about which model is best
3. Uncertainty about what to do with the output of the (best) model(s)
4. Uncertainty about whether the software works as intended
5. Uncertainty about whether the (best) model(s) hold with other data

- Bayesians use probability to describe their uncertainty in (1) and (2)
- The Bayesian approach links with decision theory, which prescribes (3)
- The Stan software does as much as we can to mitigate (4)
- By implication, other approaches / software may refer to probability but fail to handle one or more of the above five items
- These include randomization inference, Frequentist inference, supervised learning, and others

Breakout Rooms: Randomization Inference

- Pairs of people will randomly be assigned to a breakout room in Zoom
- One person in each pair should use the `sample` function in R to randomly select which person will be considered treated by passing in a character vector of size two, i.e. `sample(c("me", "you"), size = 1)`
- Compute the difference in age (in integer years) between the treated person and the control person in the pair (which can be negative) in 60 seconds

Introduction to Randomization Inference

- In the 1920s, Fisher devised a method of inference for experiments
- There are 1.264106e+14 ways to assign 25 out of 50 people to treatment, each of which would generally imply a different estimate of the Average Treatment Effect, which is computed as the mean difference
- The researcher chooses one of them AT RANDOM
- Randomization of the treatment variable creates the need for probability: Every function of the randomized treatment assignment has a probability distribution that is conditioned on everything prior to treatment assignment

Breakout Rooms: Frequentist Inference

- Everyone will be randomly assigned to a small breakout room
- When you get to the breakout room, type your age in the chat
- Exactly one person in each group should use the `mean` function in R to compute the average age in the group, i.e. `mean(c(24, 27, 12, ...))`

Introduction to Frequentist Inference

- In the 1920s, Fisher also devised a method of inference for observational data where the goal is to make an inference about a population parameter from a relatively small RANDOM sample from that population
- For example, the British population was about 36 million at the time and the government might take a sample of 500. There are $e \approx 2.7182818$ raised to the power of 6088.1808245 ways of doing so.
- The data collector chooses one way AT RANDOM and gives the data to the analyst who estimates, e.g., the population mean with the sample mean
- The probability distribution of the sample mean estimator over all the ways to draw such a sample can often be derived analytically
- In the 1930s, Neyman & Pearson devised a way of testing 2 point hypotheses about the population mean & popularized a confidence interval estimator

Quotes from *Statistical Rethinking*

- "The greatest obstacle that I encounter among students and colleagues is the tacit belief that the proper objective of statistical inference is to test null hypotheses" (p. 10)
    - Hypotheses are not models
    - Measurement error prevents deductive falsification
- "what researchers need is … a set of principles for designing, building, and refining special-purpose statistical procedures. Every major branch of statistical philosophy possesses such a unified theory. But the theory is never taught in introductory … courses. So there are benefits in rethinking statistical inference as a set of strategies, instead of a set of pre-made tools." (p. 4)

1. Bayesian data analysis
2. Multilevel models
3. Model comparison using information criteria
4. Graphical Causal Models / Directed Acyclic Graphs

Breakout Rooms: Supervised Learning

- Supervised learning departs from Frequentist statistics by not conducting null hypothesis tests, which obviates the need for probability theory
- Suppose there is a dataset of size N, which generally is neither a sample from any well-defined population nor an experiment so Fisher does not apply
- There are $\frac{N!}{n!(N-n)!}$ ways to divide this dataset into a training dataset of size n and a testing dataset of size $N - n$, where ! indicates the factorial function
- The analyst chooses one split at random (with $\frac{n}{N} \approx 0.8$), then
  - chooses parameters to minimize some loss function in the training data
  - uses the optimal parameters to predict the outcome in the testing data
  - compares the predictive accuracy to (many) other models
- Many of you will be invited to a breakout room to constitute the training data. When you get to the breakout room, type your age in the chat so that the TA can calculate the average student age.

Bayesian Inference

- Uncertainty necessitates probability to describe beliefs about unknowns
- You have beliefs about how much the S&P500 will grow by the end of 2022
- You express your beliefs with a probability distribution, such as a normal distribution with a mean of $+6\%$ and a standard deviation of 5%
- As more data comes during 2022, you update your beliefs about where the S&P500 will be at the end of 2022 to some new probability distribution
- Note the data are not, and need not be, a sample or an experiment for you to use probability distributions to describe your beliefs in a rigorous way

Perspectives on Quantitative Methodology

What is the paradigm?	What is fixed?	What is random?	What is averaged over?	What is the conclusion?
Randomization	$y_1, y_2, \ldots, y_N$	Treatment assignment	Hypothetical experiments	ATE $\neq 0$?
Frequentist	Y, $\boldsymbol{\theta}$, N	Sample inclusion	Confidence interval catches	Hypothesis test
Supervised learning	$y_1, y_2, \ldots, y_N$	Training / testing inclusion	Loss in the testing data	Some procedure predicts best
Bayesian	$y_1, y_2, \ldots, y_N, \boldsymbol{\theta}$	Beliefs about $\boldsymbol{\theta}$	Functions of posterior draws of $\boldsymbol{\theta} \mid y_1, y_2, \ldots, y_N$	Decision or action

(Dis)Advantages of Bayesian Inference

- Bayesian inference remains useful in situations other paradigms specialize in:
  - Experiments: What are your beliefs about the ATE after seeing the data?
  - Repeated designs: Bayesian estimates have correct frequentist properties
  - Predictive modeling: If you only care about predictions, use the posterior predictive distribution
- Bayesian inference is very useful when you are using the results to make a decision or take an action; other paradigms are not
- Bayesian inference is orders of magnitude more difficult for your computer because it is attempting to answer a more ambitious question
- The Bayesian approach is better suited for convincing yourself of something than convincing other people

Objectivity and Subjectivity

- Under weak and not particularly controversial assumptions, Bayesian inference is THE objective way to update your beliefs about (functions of) θ in light of new data $y_1, y_2, \ldots, y_N$
- Nevertheless, the Bayesian approach is labeled subjective because it does not say what your beliefs about θ should be BEFORE you receive $y_1, y_2, \ldots, y_N$
- Thus, if you currently believe something absurd about θ now, your beliefs about θ will merely be less absurd after updating them with $y_1, y_2, \ldots, y_N$
- The big problem is not that people believe wrong things now, but that they do not update their beliefs about θ according to Bayesian principles when they observe $y_1, y_2, \ldots, y_N$
- In fact, in some situations, observing data that contradicts people's previous beliefs makes them believe in their wrong beliefs more strongly