You Can Observe a Lot by Just Watching: Statistical Lessons from Sports

Jim Albert

September 20, 2018

Outline

- Baseball and analytics
- What is "Bayesian Statistics"? (some history)
- Two ways of thinking about learning from data
- Clever way of combining data (multilevel modeling)
- Bayesian modeling of strike/ball data (catcher framing)

A Conversation

So you've written some books about baseball ... (long pause):

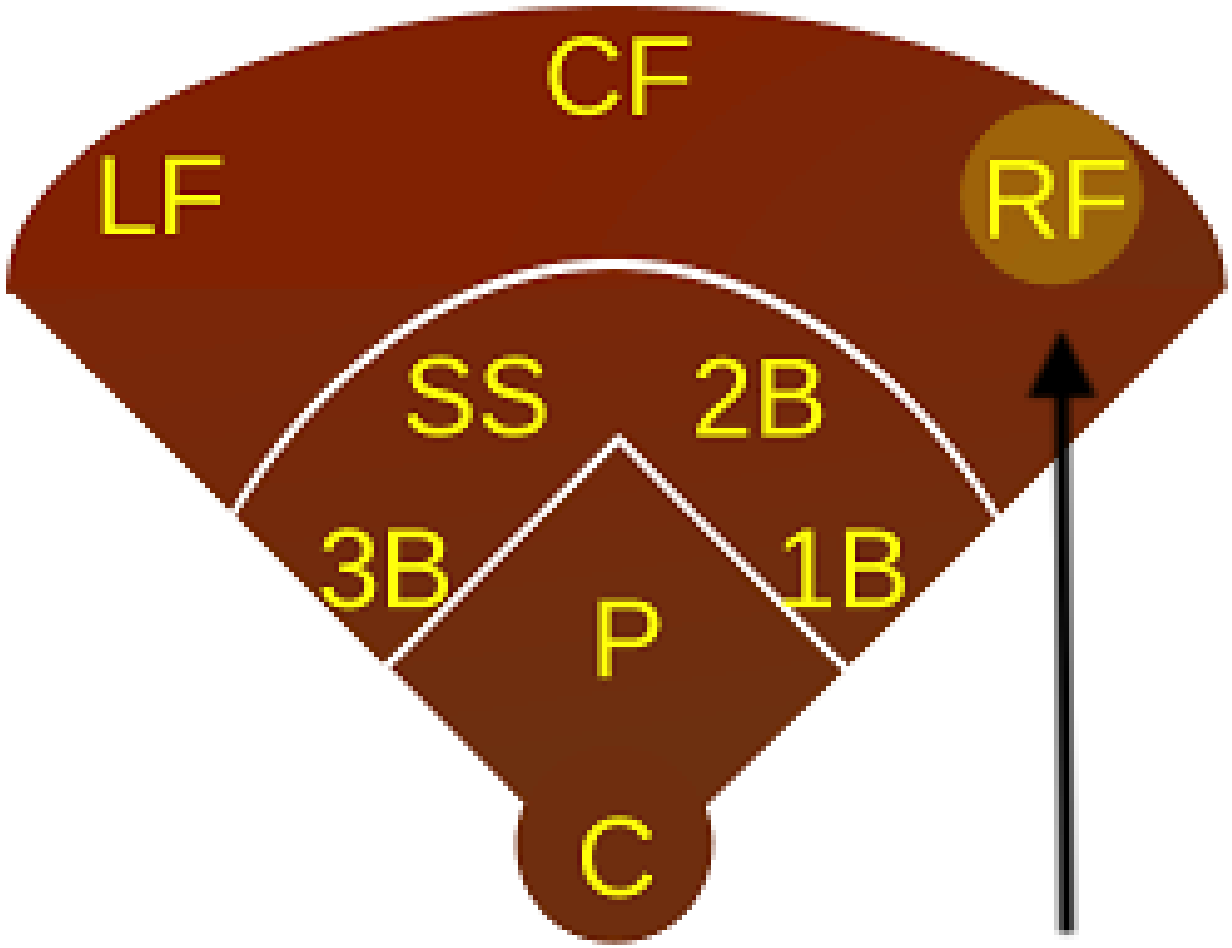
- "What do you know about baseball?"
- "What was your career batting average?"
- "How many games have you attended?"

Disclaimer

Growing up around Philly ...

- Played Little League Baseball
- Followed the Phillies
- Collected baseball cards (liked the stats on the back of the cards)
- Played baseball simulation games

What Position Did I Play?



Major League Baseball

- All 30 Major League Baseball teams have analytics groups
- What types of skills do you need to work as a data scientist?

Ad Posted by Cincinnati Reds

Duties and Responsibilities:

- Design, develop, test, implement and maintain predictive models and metrics
- Integrate new statistical analyses, models and data visualizations into existing and new applications

Knowledge, Skills and Abilities:

- Experience in a computational field, such as Statistics, Data Science, etc
- Ability to communicate effectively
- Experience with Bayesian statistics

Measuring Player Performance

- What makes a good catcher?
- How do you measure performance?
- How important is this measurement?
- Does it predict future performance?

Catcher Framing

- Ability to catch a ball so it appears to be a strike
- How do you measure this?
- How do you adjust for other variables?
- Does it matter?

John Tukey

- Famous statistician
- Book "EDA" (Exploratory Data Analysis)
- Worked on a wide range of applications

Famous quote by John Tukey

"The best thing about being a statistician is that you get to play in everyone's backyard."

My Backyard

- Gives me an opportunity to combine two passions
- Baseball
- Bayesian statistics (Way of learning about world from data)

What is Bayesian Statistics?

- Statistics is the science of learning from data
- Bayesian is a way of learning
- Not what most people learn in a stats class

Thomas Bayes?

- Born in 1701
- Presbyterian minister
- Published two papers during his lifetime
- Elected as Fellow of Royal Society

Thomas Bayes

- Studied at University of Edinburgh, Scotland
- Mathematician with deep interest in probability
- Left manuscript that was discovered by Richard Price

Bayes' Paper (1763)

LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

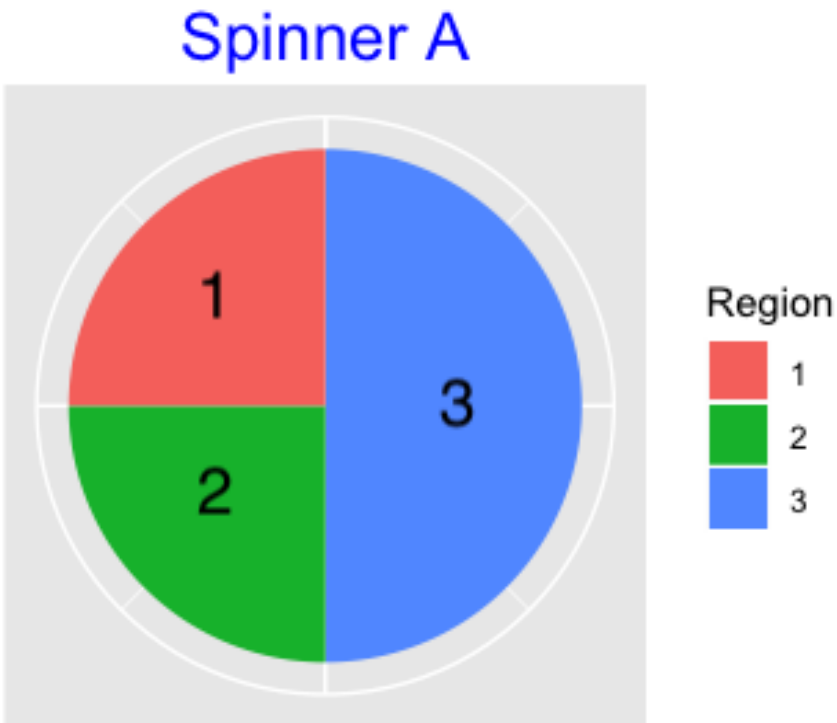
Dear Sir,

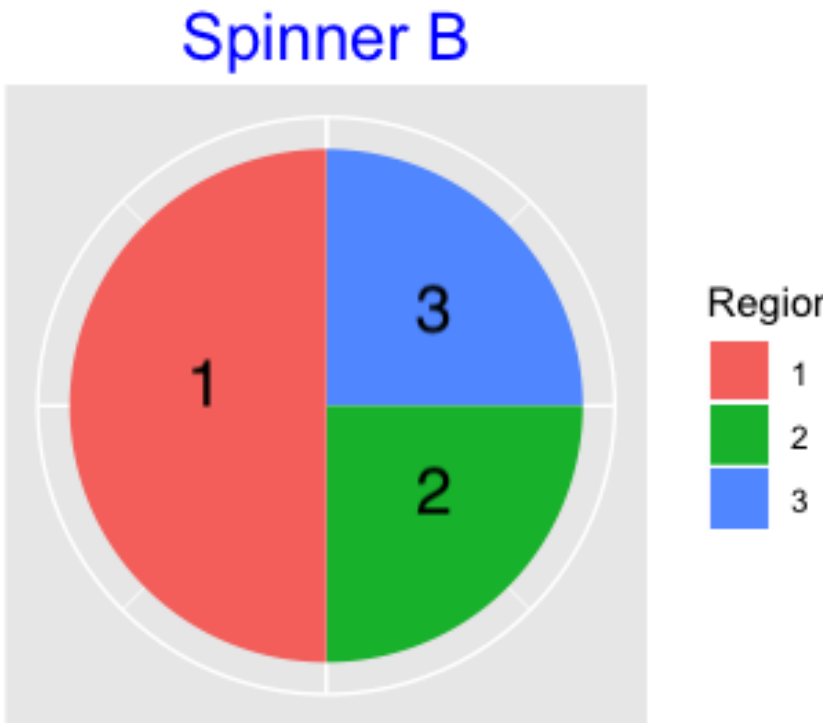
Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a com-

Bayes' Rule: Which Spinner?

- Have several spinners – Spinner A and Spinner B
- One spinner is chosen at random and is spun.
- We observe "1"
- What is the chance Spinner A was spun?

Two Possible Spinners





Bayes' Rule

Posterior Odds of Spinner A is equal to the product:

- Prior Odds of Spinner A
- The Likelihood Ratio

$$LR = \frac{Chance\ of\ Data\ from\ Spinner\ A}{Chance\ of\ Data\ from\ Spinner\ B}$$

That is,

$$Post\ Odds = Prior\ Odds \times LR$$

Step 1: The Prior

- I don't know which spinner is chosen
- No reason to favor one or the other
- So

$$\textit{Prior Odds} = 1$$

Step 2: Observe Data

- Suppose the spinner spin is "1"
- Likelihood Ratio is

$$LR = \frac{\textit{Chance of "1" from Spinner A}}{\textit{Chance of "1" from Spinner B}}$$

- So

$$LR = \frac{1/4}{1/2} = \frac{1}{2}$$

Step 3: Update (the Posterior)

- Posterior odds is

$$\begin{aligned} \textit{Post Odds} &= \textit{Prior Odds} \times \textit{LR} \\ &= 1 \times \frac{1}{2} = \frac{1}{2} \end{aligned}$$

- More likely Spinner B was spun

Bayes' Rule: Testing for Breast Cancer

- Woman takes a mammogram – result is positive – does she have breast cancer?
- Prior – chance of cancer based on age, family history, etc.
- Data – result of the mammogram test
- Posterior – revise one's opinion about cancer based on the test result

Bayes' Rule: Spam Filtering

- You receive an email message – the header contains the word "cash" – is it spam?
- Prior belief that the message is spam
- Data: see "cash" in the header
- Posterior: it is more likely that the message is spam

Bayes' Rule: Who wrote "In My Life"?

› News › Science

Sir Paul McCartney 'misremembers' writing 'In My Life' – it was really John Lennon, says Harvard analysis

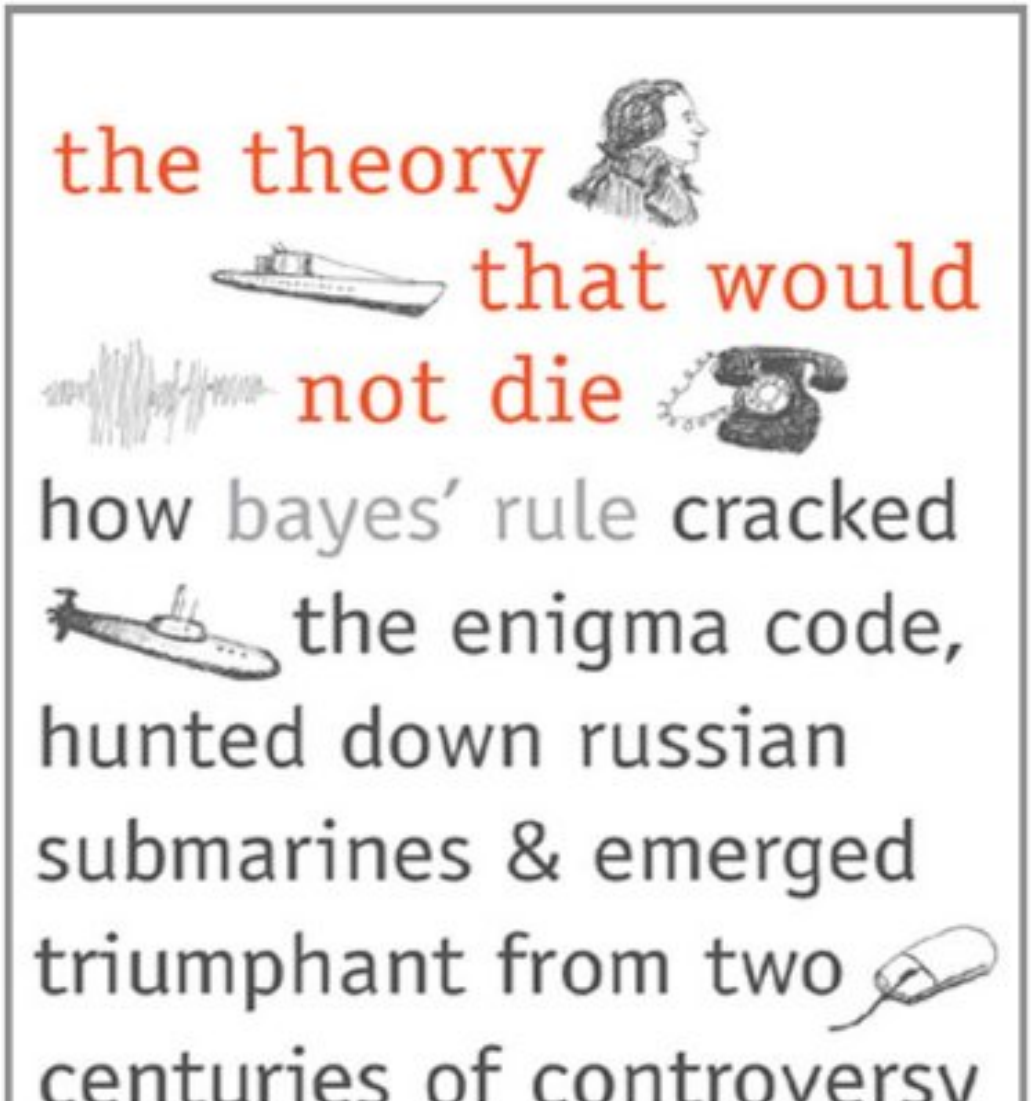
share

Save 29

Who wrote "In My Life"?

- Either it was John Lennon or Paul McCartney
- Prior – give each Beatle a probability of $1/2$
- Data – chord transitions of the song
- Posterior – Most likely John wrote the song

History of Bayes Rule: Book by Sharon Bertsch McGrayne



History of Bayes' Rule

1. Maybe it should be called Laplace's Rule
2. In the 20th century, another way of learning from data became more popular
3. Bayes' rule was heavily used during World War II (cracking the German code, finding Russian submarines), but this work was hidden

Pierre Simon Laplace: 1749-1827

- One of the greatest scientists of all time
- Made fundamental developments of mathematics, statistics, physics and astronomy

Laplace and Bayes' Rule

- Laplace independently discovers Bayes' rule
- Spent 40 years clarifying, simplifying, generalizing the result
- Interestingly, he discarded Bayes' rule for a "frequency approach" to learning from data

20th Century: R.A. Fisher
Father of Modern "Frequentist" Statistics

- British statistician and geneticist
- Made fundamental contributions to Statistics
- First to use "Bayesian", but was a fierce opponent of Bayes

Two Ways of Thinking About Probability

- Flipping a coin – Is it fair or biased?
- 20 flips and get 15 heads
- What do you conclude?

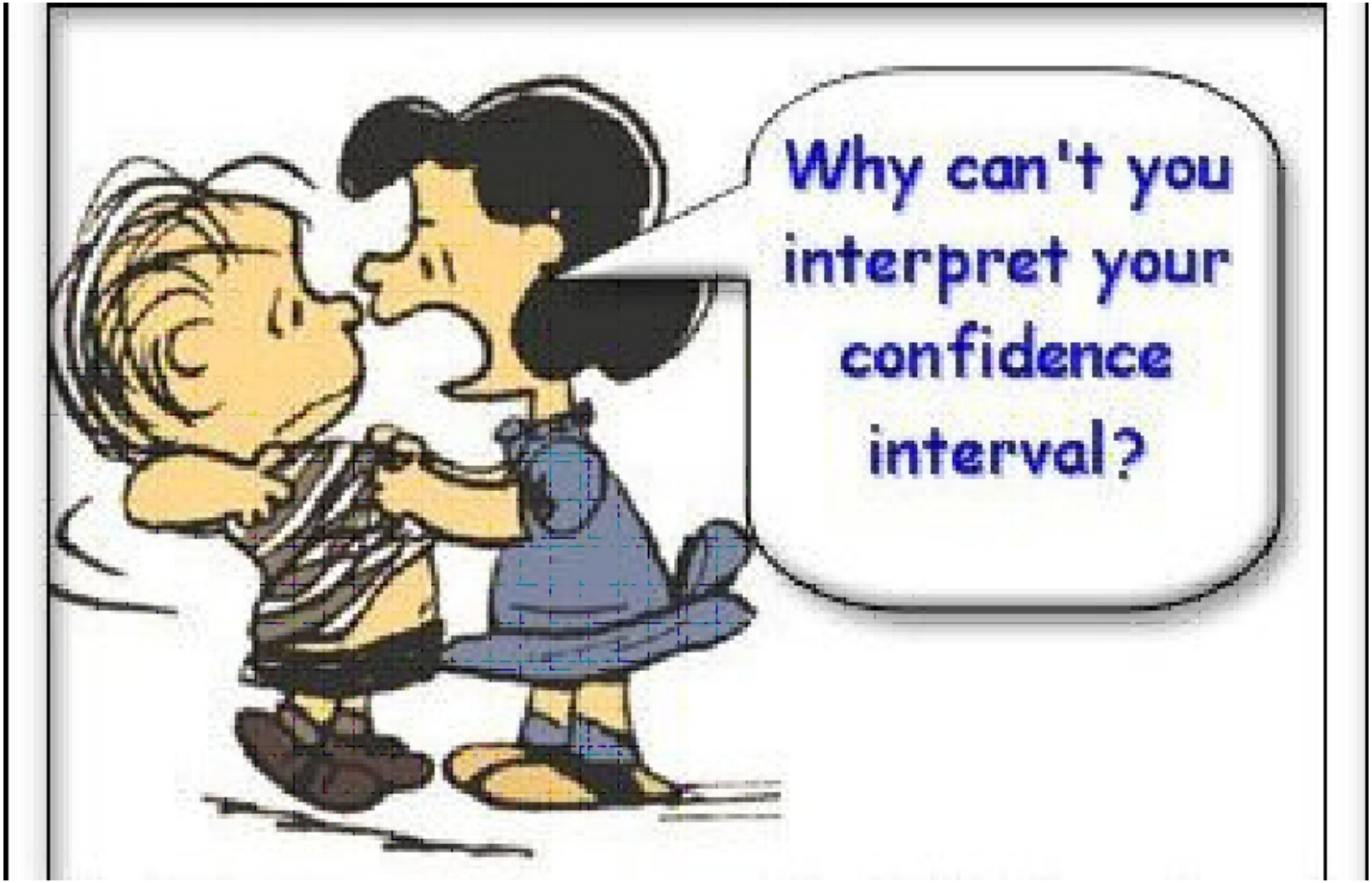
Frequentist Viewpoint

- Think of data as random
- What's the chance of getting 15 or more heads if the coin is fair?
- If this probability is very small, reject that coin is fair

Bayesian Viewpoint

- Think of state of coin as random
- Undecided – P(fair coin) = P(biased coin)
- If you get 15 heads, what is the chance coin is fair?

Frequentist Methods Harder to Interpret



The 20th Century

- R. A. Fisher made some fundamental contributions to statistics
- Mathematical elegant, easy to compute
- Frequentist thinking became the foundation
- Today statistics is taught from a frequentist perspective

McGrayne: Bayes' Rule Faces a Crisis

- Was this 18th century theory doomed to oblivion?
- Due to Fisher and others, the world favors the frequentist theory
- Big hurdle in implementing Bayes' Rule: Computation

Bayes Rule Requires One to Integrate

- To implement Bayes' rule, need to compute high-dimensional integrals

$$p(x) = \int ... \int g(\theta_1, ..., \theta_p) d\theta_1 ... d\theta_p$$

- Can't directly integrate, numerical approximations are expensive
- Is Bayesian inference possible?

Markov Chain Monte Carlo to the rescue (1990)

- In 1990, statisticians discovered algorithms for simulating from Bayesian posterior distributions
- Randomly stepping through probability distributions
- It become possible to implement Bayes' rule to fit large statistical models

Bayesian Statistics Today

- More efficient simulation algorithms for fitting Bayesian models
- Faster and faster computers
- Computational burden does not exist
- Bayesians can easily fit multilevel models

What's a Multilevel Model?

- Clever way of combining data from different sources
- Famous statistical paradox that motivates this type of model

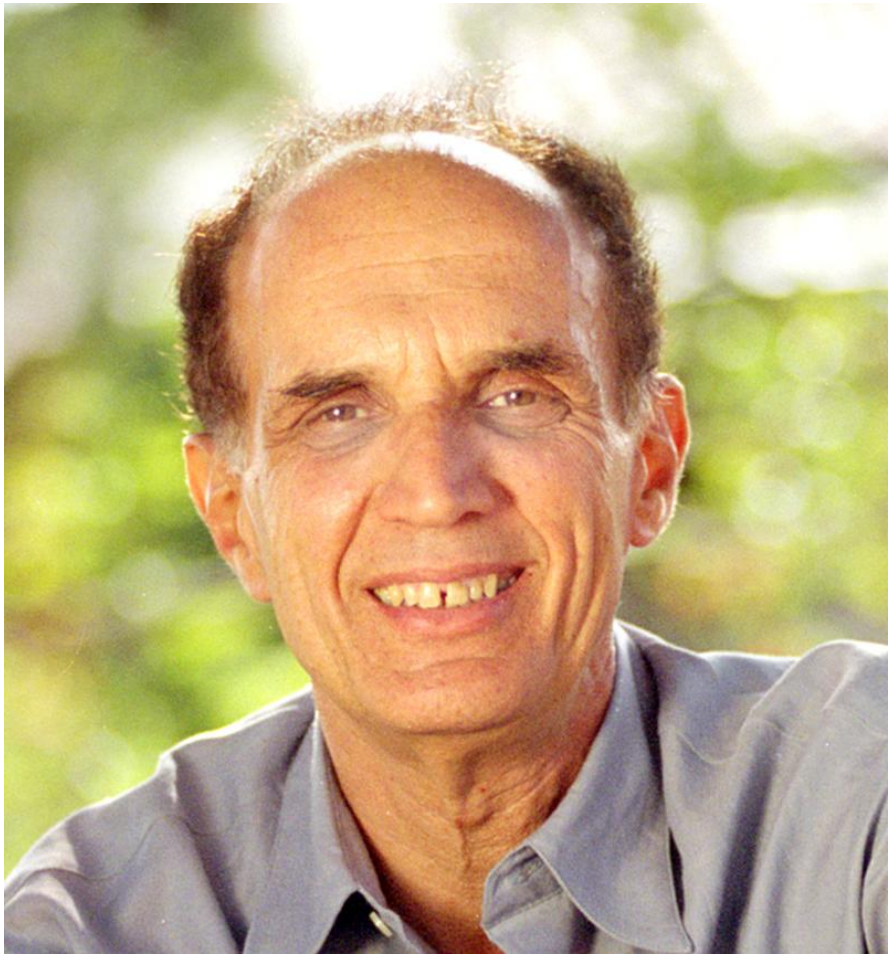
1955: Stein's Paradox

- Suppose you want to learn about the "true" mean weight of Hershey's candy bars.
- Take a sample of 20 bars – find the average of the weights.
- This can be shown to be the "best" guess at the true mean

Estimating Three Means

- Suppose you want to learn about ...
  - the true mean weight of Hershey's candy bars
  - the true mean weight of Reese's peanut butter cups
  - the true mean weight of packs of Twizzlers
- Samples of each candy.
- Stein's paradox – the individual average weights are not the best estimate.

Brad Efron and Carl Morris



Making Sense of Stein's Paradox

- Efron and Morris wrote papers in the 1970's explaining how Stein's Paradox works
- Wanted to explain to a general audience
- Needed a simple illustration of the "better" estimate
- Looked at batting averages in baseball

Scientific American Paper

Stein's Paradox in Statistics

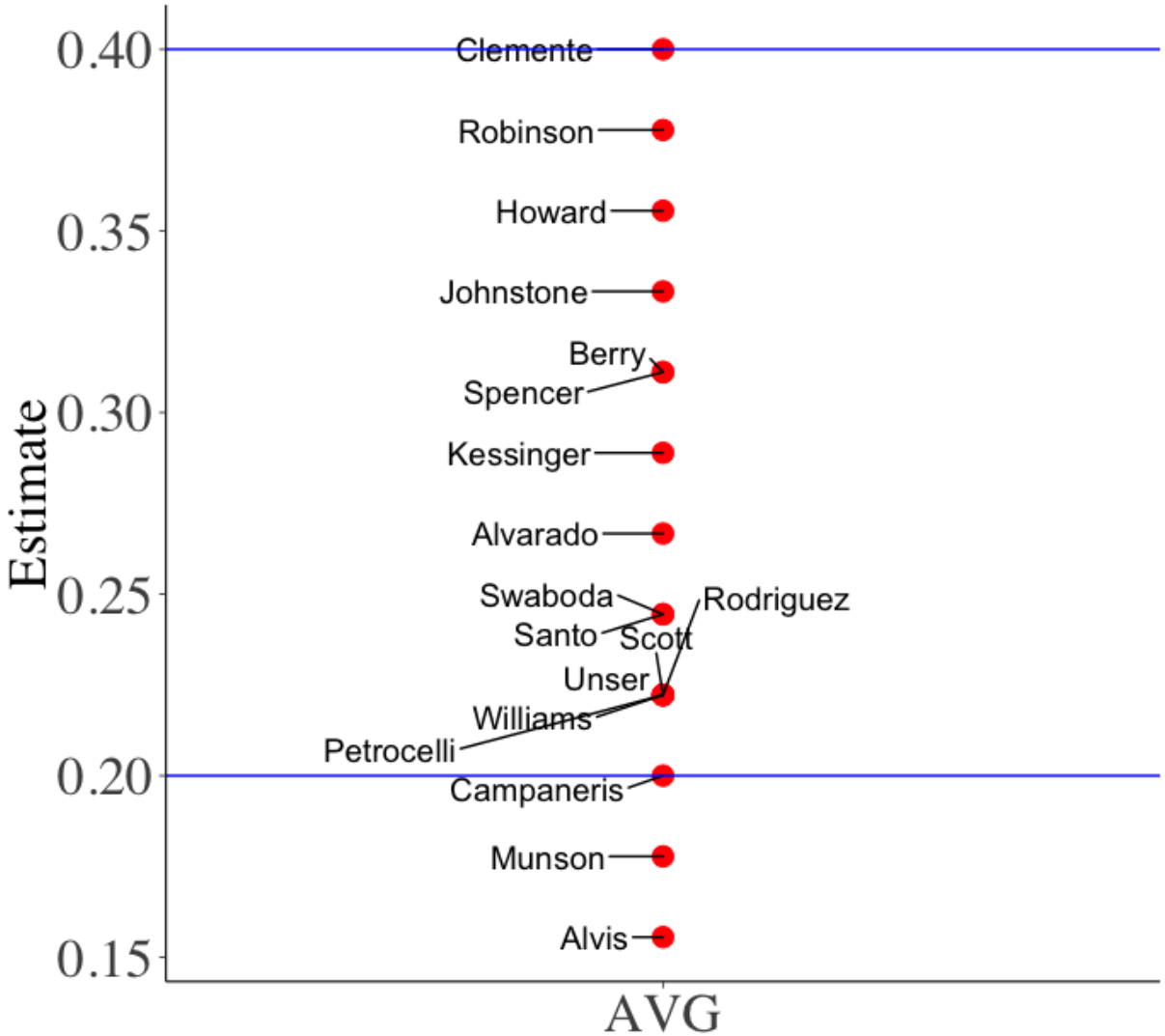
The best guess about the future is usually obtained by computing the average of past events. Stein's paradox defines circumstances in which there are estimators better than the arithmetic average

by Bradley Efron and Carl Morris

Efron and Morris Got Baseball Data from Sunday Paper (1970)

- Looked at back page of the sports section
- Roberto Clemente had 18 hits in 45 at-bats
- Collected batting averages for 18 players who had exactly 45 at-bats
- Want to predict batting averages at end of season
- Current AVGs are poor predictions

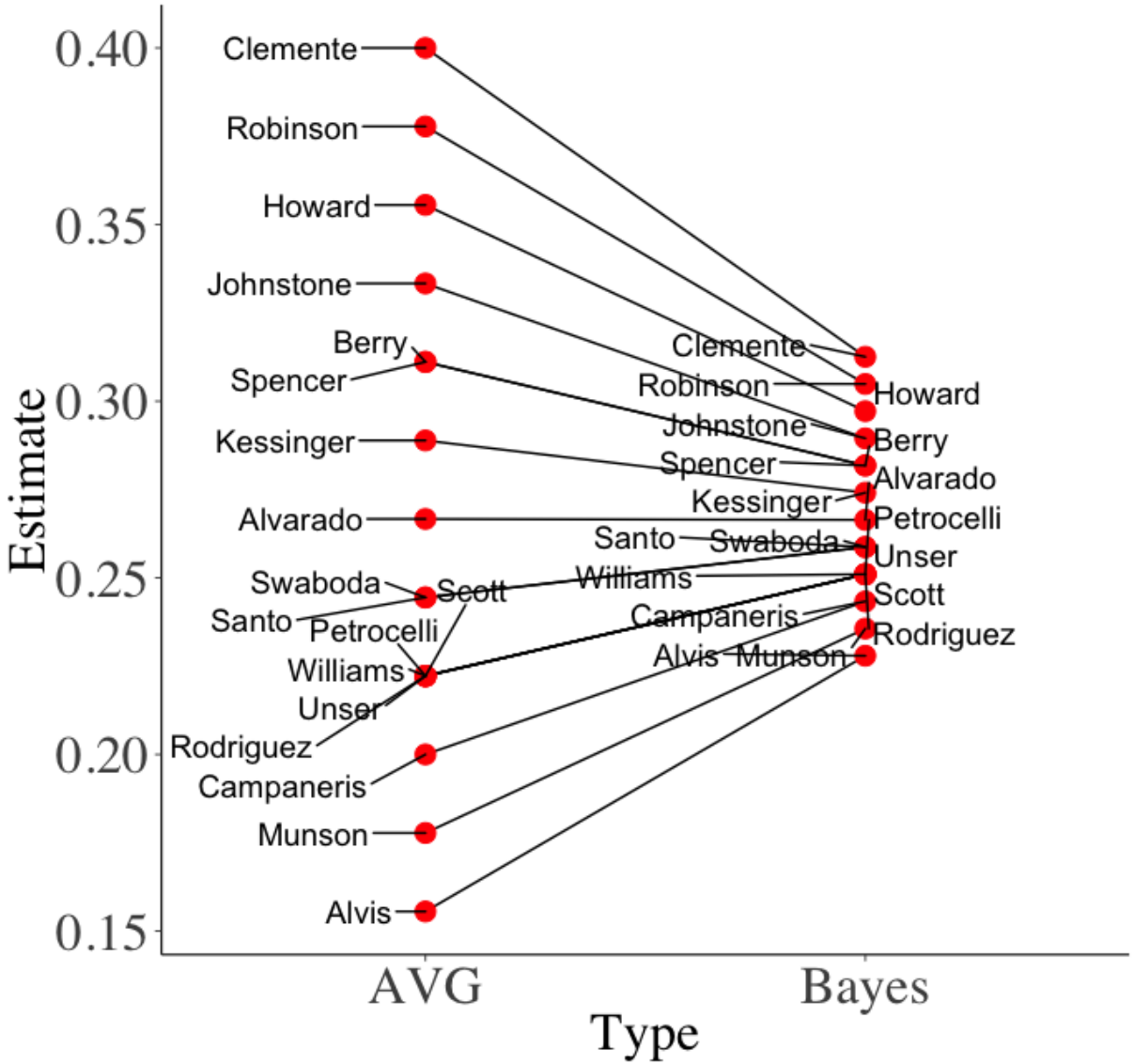
Graph of Current AVGs



Apply a Bayes Multilevel Model

- Interested in learning about player's "true" hitting probabilities
- Bayesian prior assumes these probabilities come from a common distribution
- Believe these probabilities are similar in size
- Estimate probabilities from posterior distribution

Bayes Adjust AVGs Towards Mean



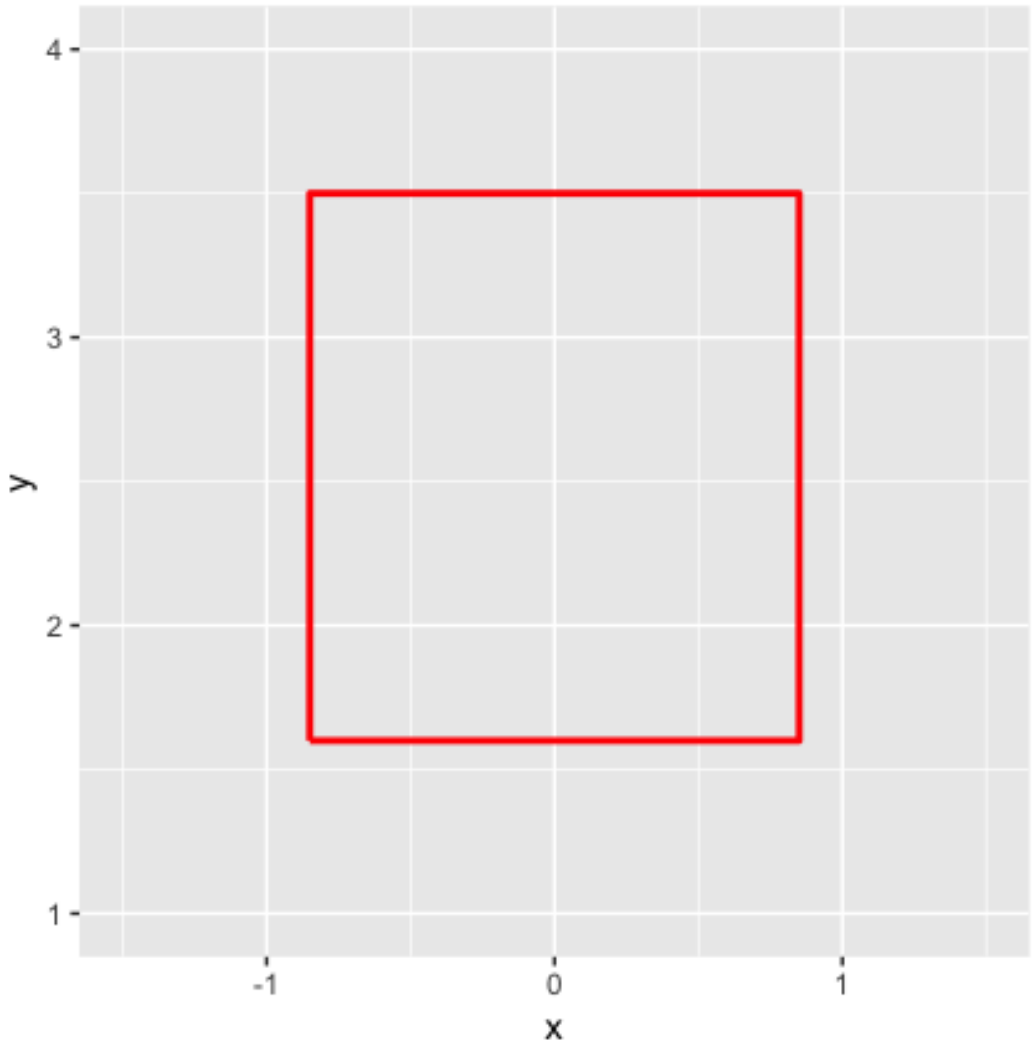
Multilevel Modeling Works in Many Situations

- Combining achievement test scores from many schools (small and large).
- Predicting election results (Nate Silver)
- Baseball applications
- Estimating catcher framing

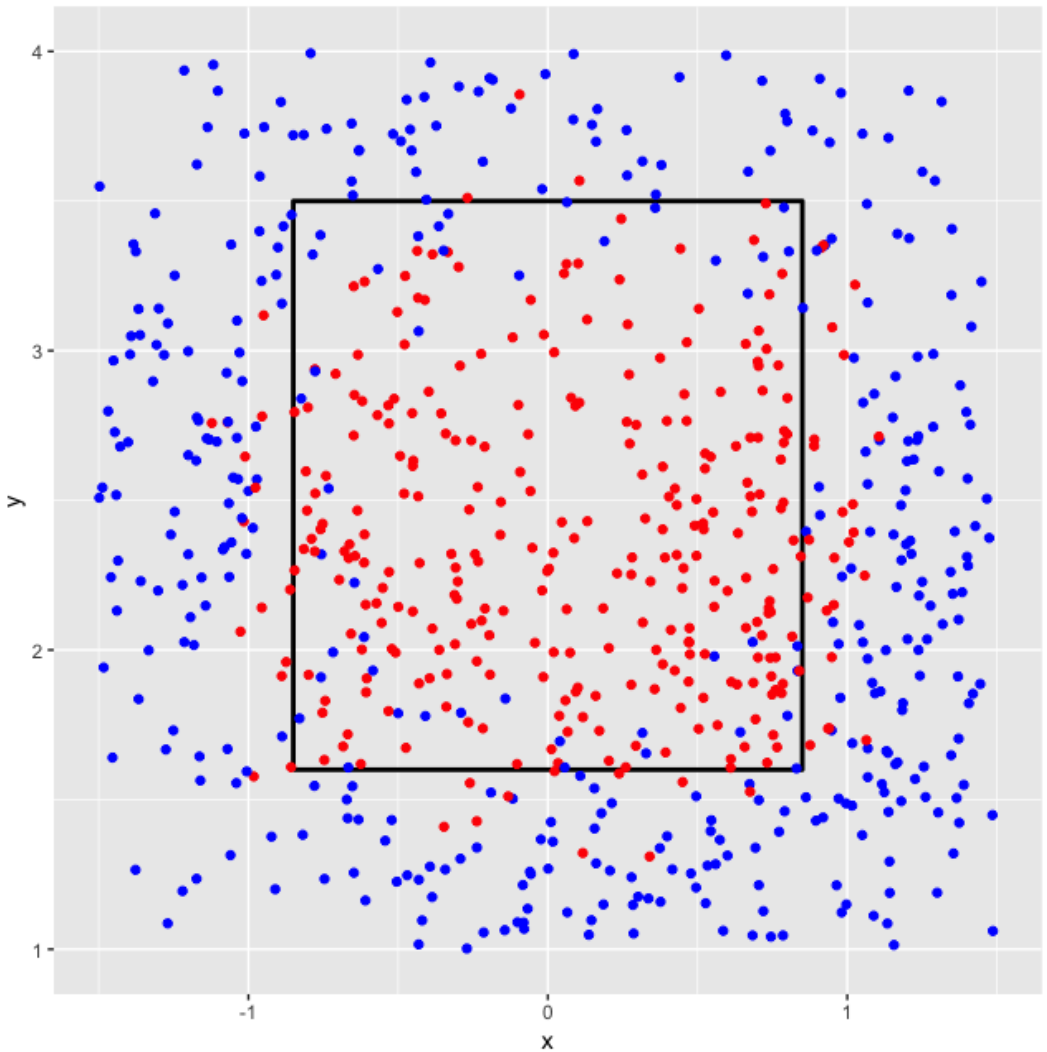
Called Balls and Strikes

- Pitches are thrown towards a "strike zone"
- Pitches are called "strikes" or "balls"
- Pitches inside zone should be strikes

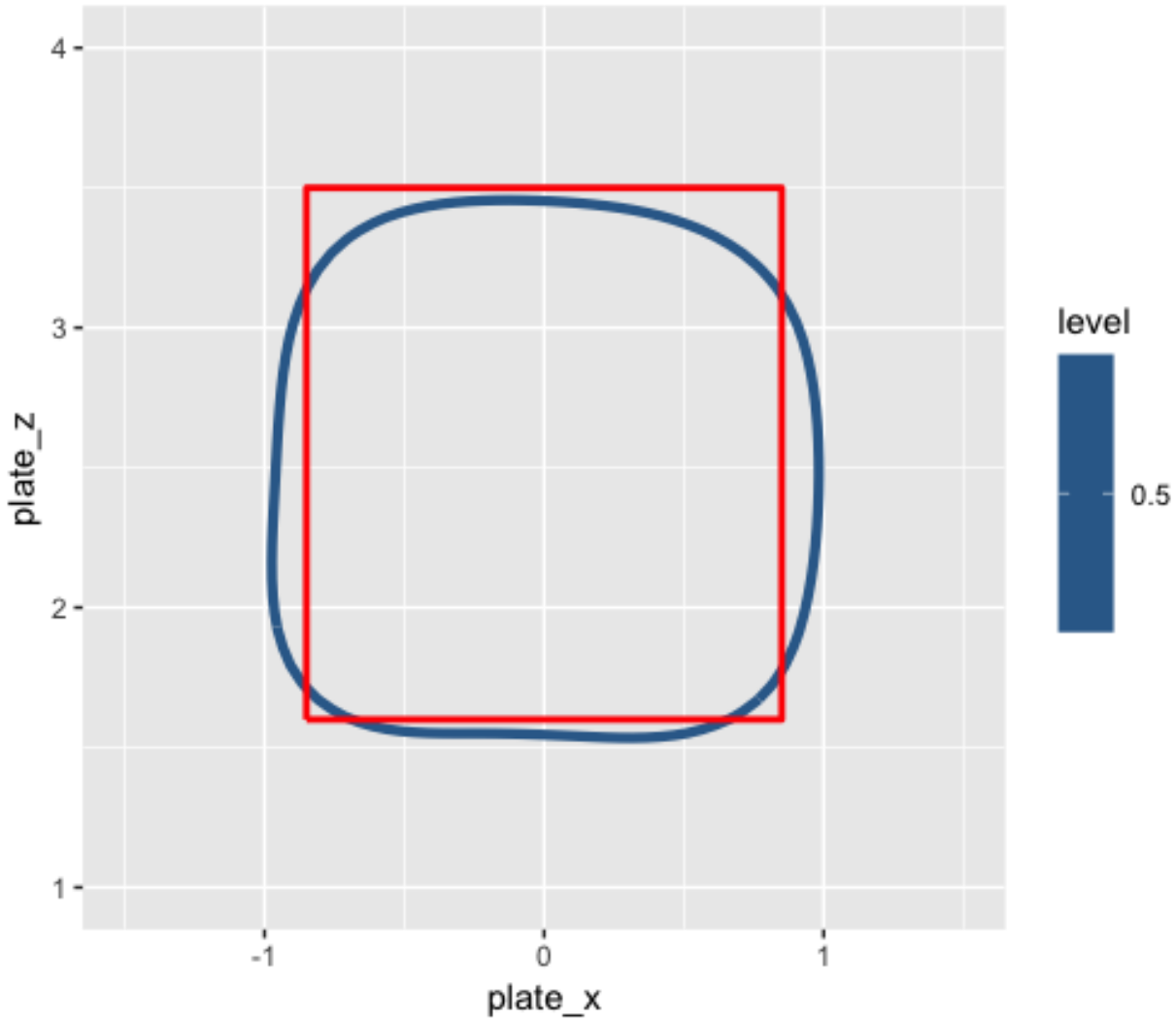
Strike Zone



Results of 1000 Called Pitches (Red = Strike, Blue = Ball)



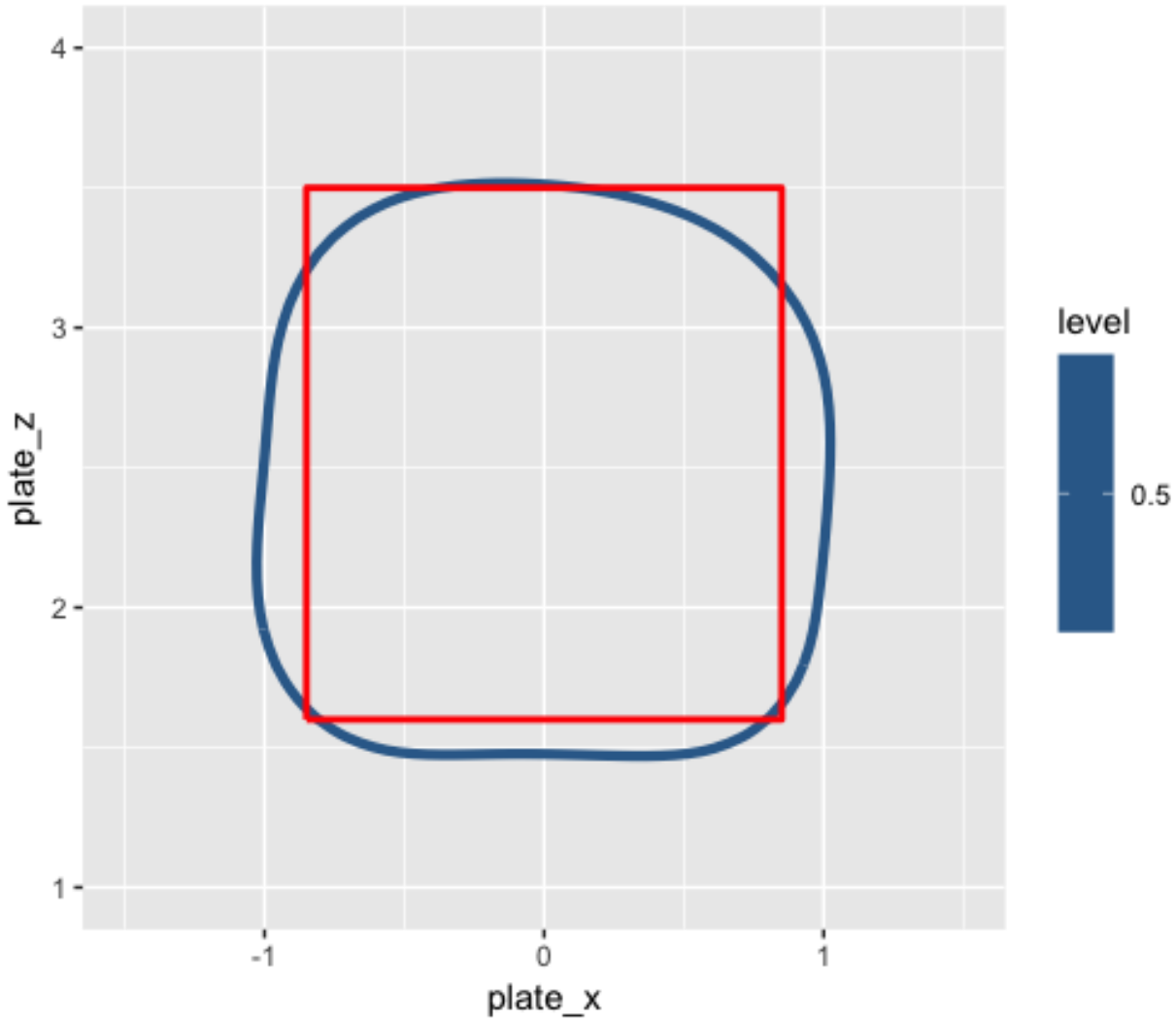
Actual Strike Zone



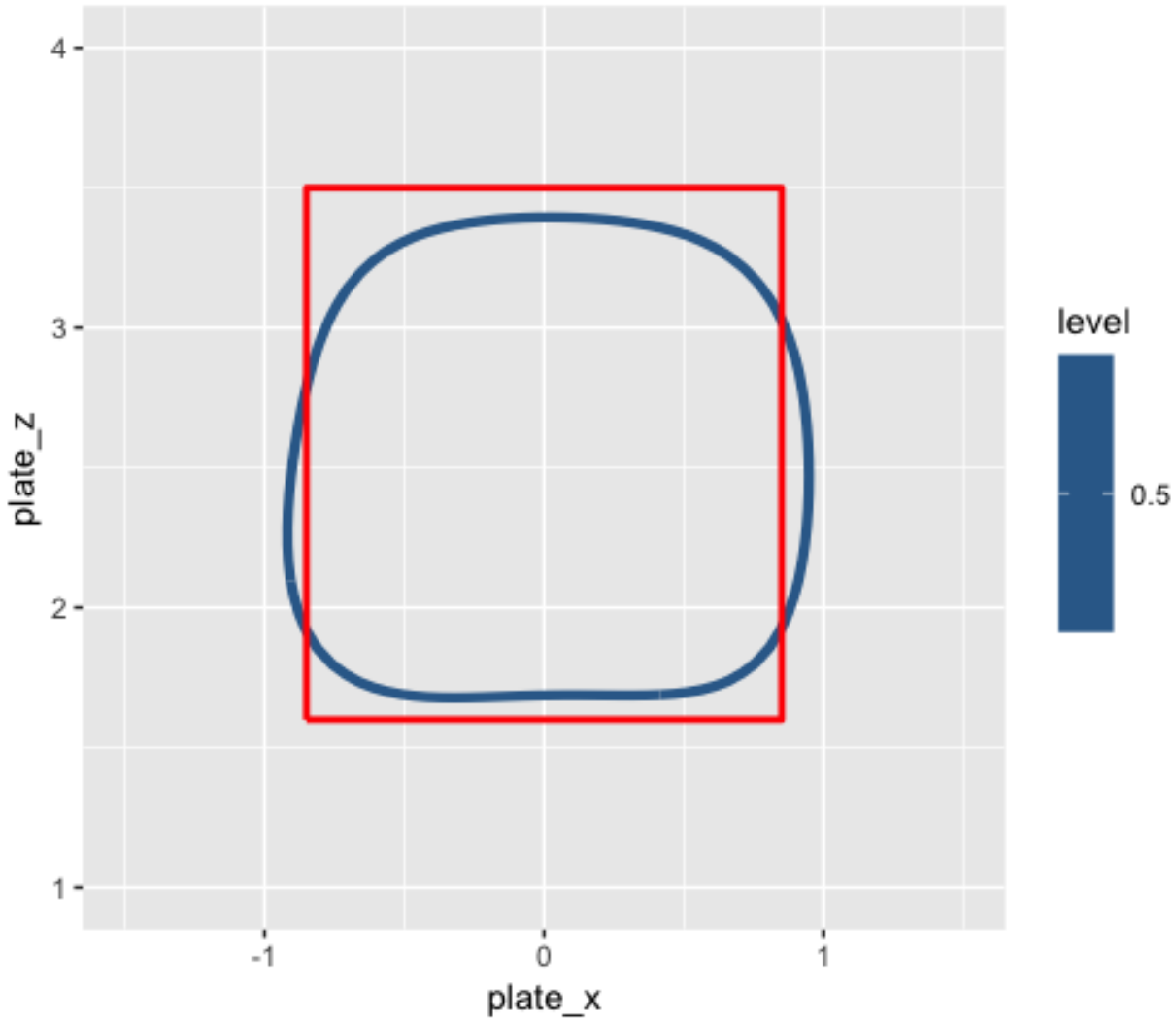
What influences balls and strikes?

- The umpire
- The pitcher and batter
- The count
- The catcher

Actual Strike Zone - 2-0 Count



Actual Strike Zone - 0-2 Count



Catcher Framing

- Catcher can influence the called pitch
- Subtle way the ball is caught
- How do you measure it?
- How big an effect is it?

Fit a Bayesian Multilevel Model

- Outcome – called pitch (strike or ball)
- Inputs:
  1. Location
  2. Pitcher effect
  3. Batter effect
  4. Umpire effect
  5. Catcher effect
- We are measuring catcher framing adjusting for all over variables

A Good Catcher Framer

- Gives the defensive team more called strikes (instead of balls)
- Each called strike saves about 0.03 runs per called strike
- Best framers save 10-20 runs scored for their teams

A Related Issue: Situational Effects

- We like to compute situational stats
- How a player performs home and away, against different pitchers, in clutch situations
- See interesting variation (some players have clutch performance)
- Do these mean anything? (Are there players with clutch abilities?)

Sports Analytics: The Future

- More and more data
- Have the opportunity to learn more about player performance
- Goal to separate out the signal (ability) from the noise (performance)
- Bayesian lens give us a natural way to learn about player ability