Welcome to the Bayes Jam

QSC 2021

Nick Klagge, Andrew Nguyen

FRB NY, Board of Governors

Disclaimer

- We started this project to learn something about Bayesian analysis with a fun topic
- We are by no means experts!
- If the Federal Reserve System has views on the NBA, this does not represent them

Intro

- NBA playoffs are played in 'best of seven' series
  - Teams are matched up by "seed" or rank - 1 plays 8, 2 plays 7, and so on
  - Each series is best-of-seven. First team to win four games advances
  - Higher seeded team gets "home-court advantage"
  - Standard (mostly) pattern of H-H-A-A-H-A-H
- Some questions
  - Is there any game that is particularly important for winning a series?
  - How important is home-court advantage?

Data

- The outcome of every playoff game from 1998-2020 from basketballreference.com, scraped with Python
- Elo rating data from 538, from R *fivethirtyeightdata* package
  - A relative ranking system updated after every game, originally developed for chess. Can be used to infer the probability of winning a single game

> " Two players with equal ratings who play against each other are expected to score an equal number of wins. A player whose rating is 100 points greater than their opponent's is expected to score 64%; if the difference is 200 points, then the expected score for the stronger player is 76%."

Analysis framework

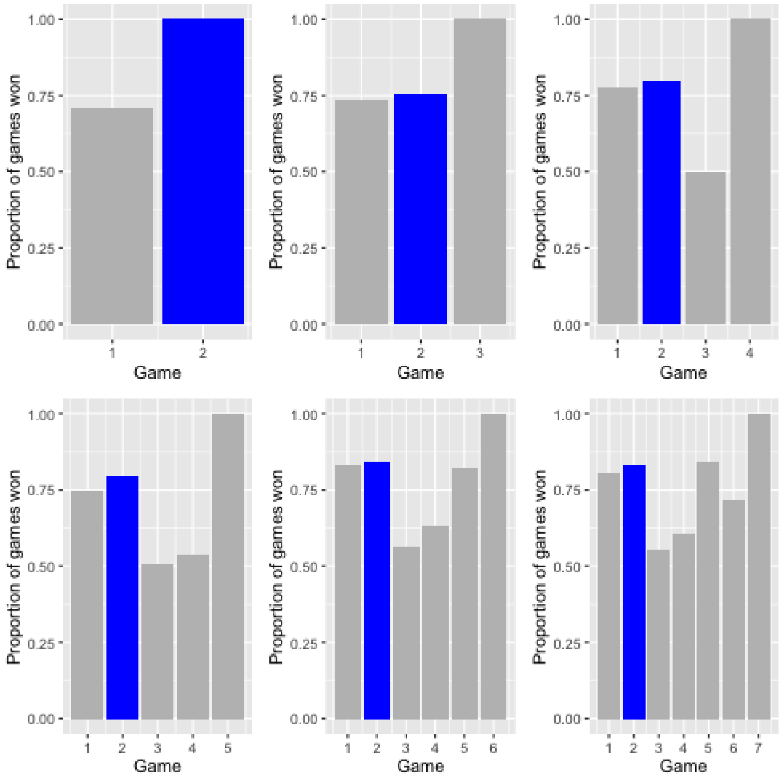
- Analyze all playoff series from the perspective of the higher seeded team
  - Want to be consistent when considering which games matter, given standardized home-away pattern
- Examine Elo, home court advantage, individual game influence
- We'll look at summary statistics, logit regression, and Bayesian estimation
- We focus most on games 1-4, as these are played in every series
  - Analyzing games 5-7 is tricky because their existence is conditional on earlier game outcomes
  - Games 1-4 have a fully consistent home/away pattern, while games 5 and 6 have differed at times

High Level facts

- Elo implies that the higher seeded team should win individual games 58% of the time on average
- In fact, they win about 62% of games
- This is largely explained by home court advantage: win frequency varies a lot by game number
- Higher seeded teams win about 75% of series ("best-of" favors the better team)

game_number	away_win_pcnt	home_win_pcnt
1	—	0.73
2	—	0.75
3	0.46	—
4	0.49	—
5	0.43	0.78
6	0.50	0.56
7	—	0.73

Game 2 outcome most highly correlated with subsequent game outcomes



Game 2 outcome most highly correlated with winning the whole series

term	estimate	std.error	p.value
(Intercept)	−2.87	0.49	0.00
win_game_1TRUE	2.28	0.39	0.00
win_game_2TRUE	2.52	0.40	0.00
win_game_3TRUE	2.35	0.42	0.00

term	estimate	std.error	p.value
(Intercept)	−5.35	1.26	0.00
win_game_1TRUE	2.23	0.40	0.00
win_game_2TRUE	2.32	0.41	0.00
win_game_3TRUE	2.27	0.42	0.00
p_win_game_1	4.70	2.11	0.03

Home court advantage still matters if we account for Elo

term	estimate	std.error	statistic	p.value
(Intercept)	−0.39	0.09	−4.33	0.00
elo_diff	0.01	0.00	6.47	0.00
is_home_teamTRUE	1.19	0.11	10.89	0.00

"I know I am just an algorithm... No one knows who I am, or what I do. But that all changes today."

AI G Rhythm on Bayesian updating

A Bayesian perspective

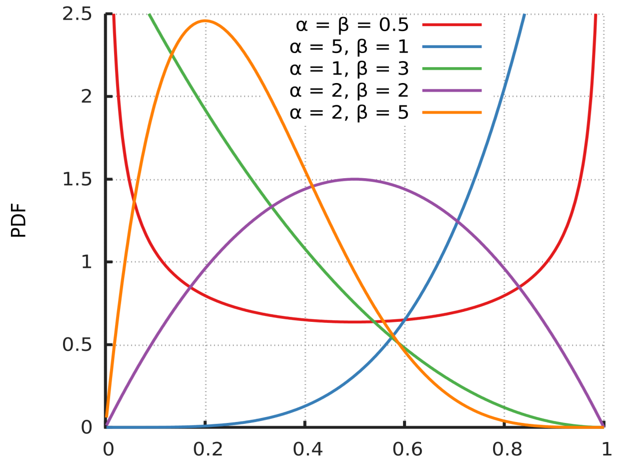
- Bayesian reasoning focuses on taking existing probability estimates and updating them based on new information
- This seems like a natural framework for our questions: we're interested in how the serial outcomes of individual games influence the probability of an overall series win
- Let's start by assuming we don't know anything about the two teams playing in a given series besides which one is higher-seeded. How estimate of their series win probability?
    - In a "frequentist" perspective, we could estimate their win probability by the sample average series win frequency for higher-seeded teams: 75%
    - In reality, we know that there's not just a single answer to this question: some series are very closely matched and others are not
    - That 75% sample average is made up of a range of higher and lower individual win probabilities
    - Bayesian reasoning lets us represent our estimate of the win probability with a probability distribution

The beta distribution

We let the higher-seeded team's series win probability p be described by a beta distribution. why?

- Bounded between 0,1
- Extremely flexible with only two parameters
- Estimate α and β (more on these in a second)

$$\mathrm{E}[X] = \frac{\alpha}{\alpha + \beta}$$



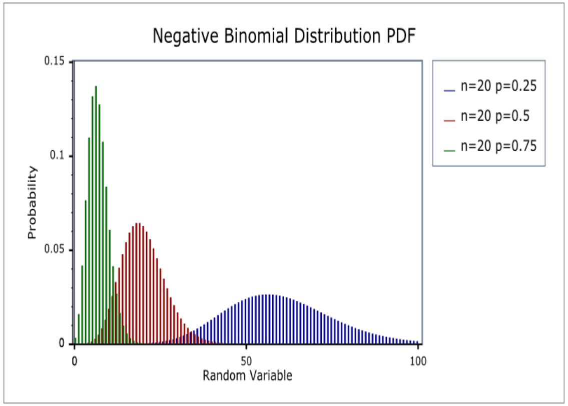
Developing a prior

- What is our "prior" estimate of the distribution of win probability, before observing any game outcomes?
- The most "uninformed" prior would be a beta(1,1), which is equivalent to the uniform distribution
- But we definitely know at least a little more than that: at minimum, we know the home team is more likely to win (both because they are higher ranked and because of home court advantage)
- We can use the pre-series Elo ratings and the negative binomial distribution to develop our prior

The negative binomial distribution

The negative binomial distribution is based on an experiment satisfying the following conditions:

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in either a success or a failure.
3. The probability of success p is constant from trial to trial
4. The experiment continues until a total of r successes have been observed



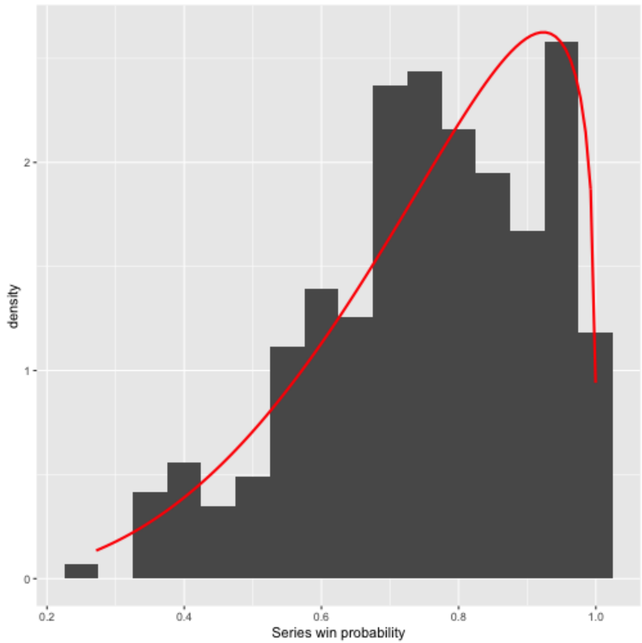
This sounds a lot like the conditions of a best-of-seven series!

Negative binomial series win probabilities

- Given any two teams' pre-series Elo ratings, we can calculate the game-wise win probability, and feed this to the negative binomial distribution to estimate the series-wise win probability.
- We know that the assumptions of the negative binomial don't fully hold in the NBA playoffs: home-court advantage means that the win probability varies in different game numbers. But maybe it's close enough to develop a reasonable prior
- We estimated the win probabilities for each series in the dataset using Elo and the negative binomial, then fit a beta distribution to that sample
- We found that the mean of this distribution was slightly lower than the empirical win frequency for the higher seeded team
    - This makes sense because the negative binomial does not account for home court advantage
    - We adjust the game-level win probability upward to approximate this. We find that a 9% increase in the higher-seeded team's game-level win probability results in a distribution with a mean that matches the empirical frequency.

Our prior on series win probability

```
beta_fit <- MASS::fitdistr(po_dat_wide$prior_p_series_win, dbeta,
                    start = list(shape1 = 1, shape2 = 1))
# beta estimates: 3.86, 1.24
alpha0 <- beta_fit$estimate[1]
beta0 <- beta_fit$estimate[2]
# mean p(series_win) is 76%
alpha0/(alpha0+beta0)
```



Updating the prior

- The beta distribution has the special property of being a *conjugate prior* for Bernoulli trials
- This means that, when we have a beta-distributed prior on a binomial probability, then observe a sample of Bernoulli trials, the posterior distribution after we update on the outcomes is also a beta distribution

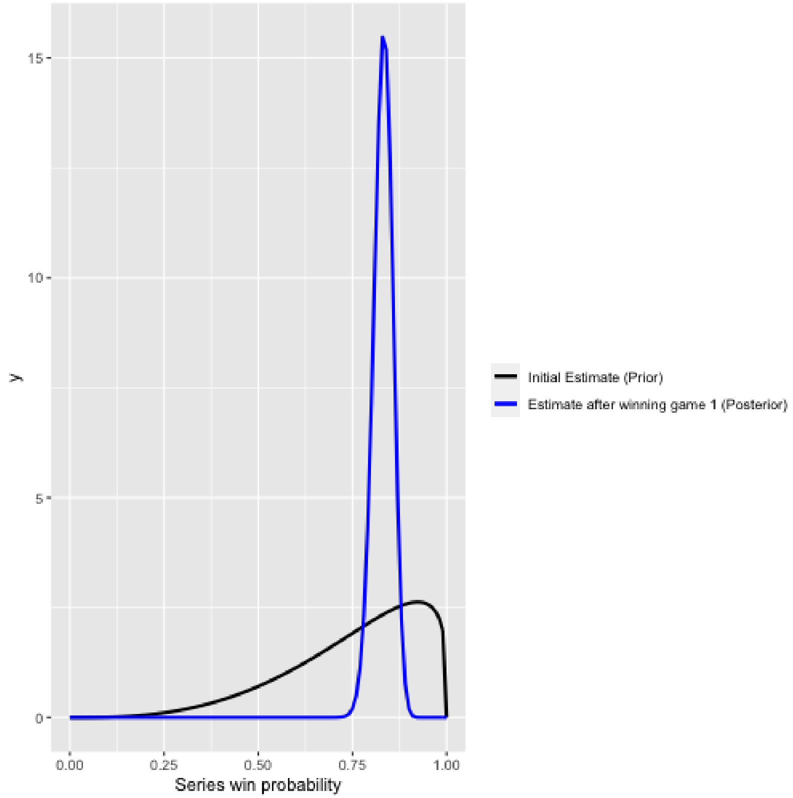
In particular, when we observe new successes and failures, the new Beta distribution becomes

$$Beta(\alpha_0 + successes, \beta_0 + failures)$$

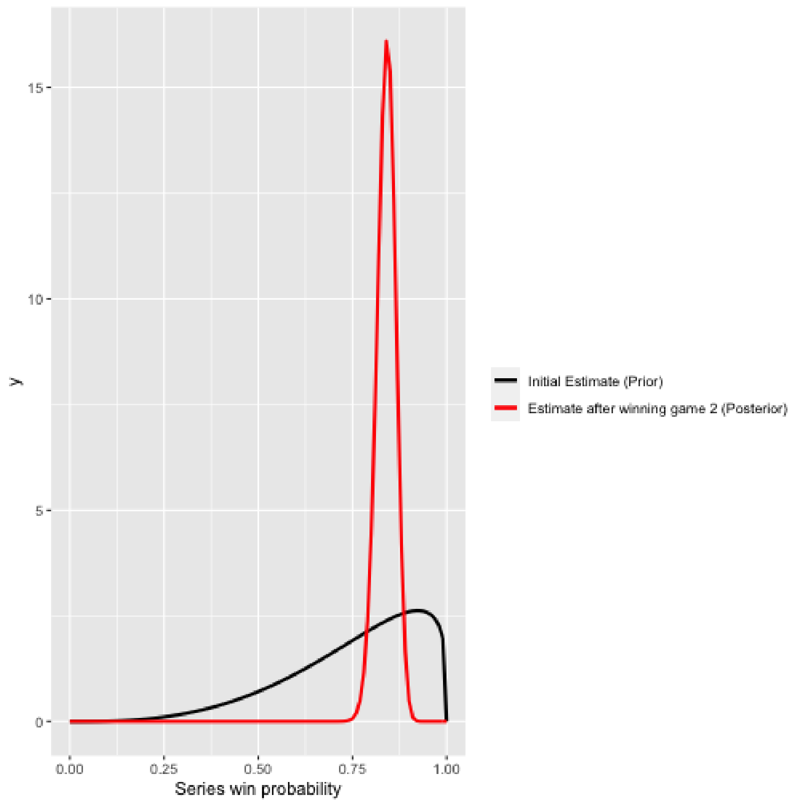
In a basketball playoff context, say we observe n series and in m of those, the teams won game 1 and the series. We can update our prior accordingly

$$Beta(\alpha_0 + m, \beta_0 + (n - m))$$

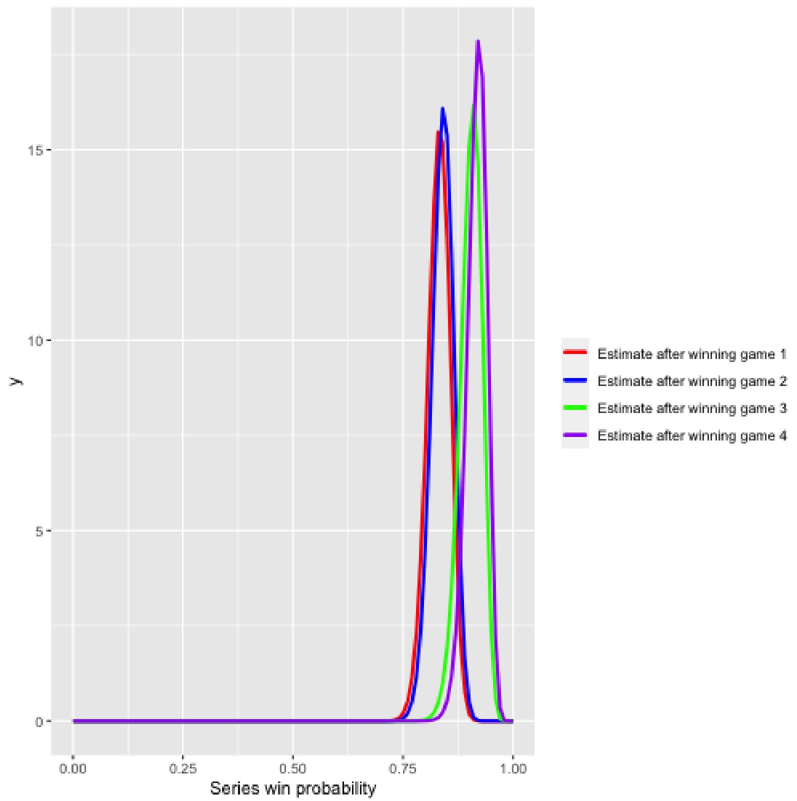
Update conditional on game 1 win



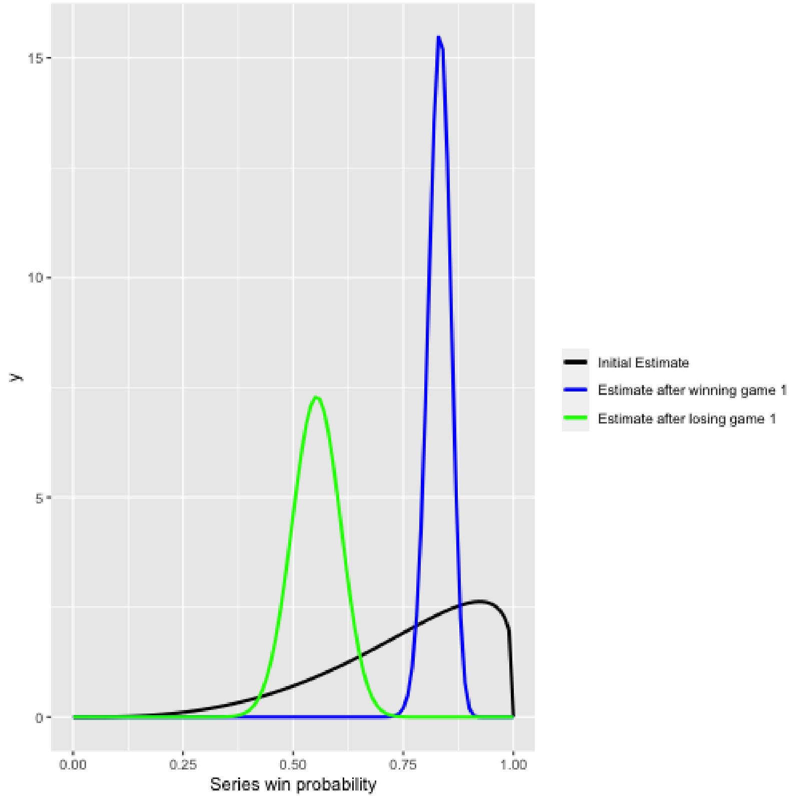
Update conditional on game 2 win



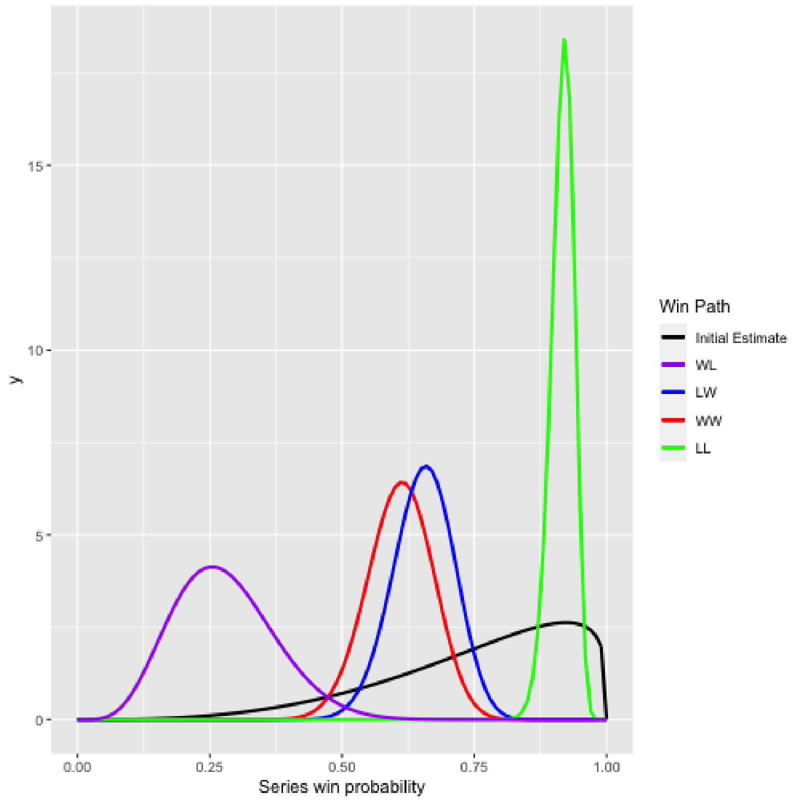
Update conditional on winning each of first four games



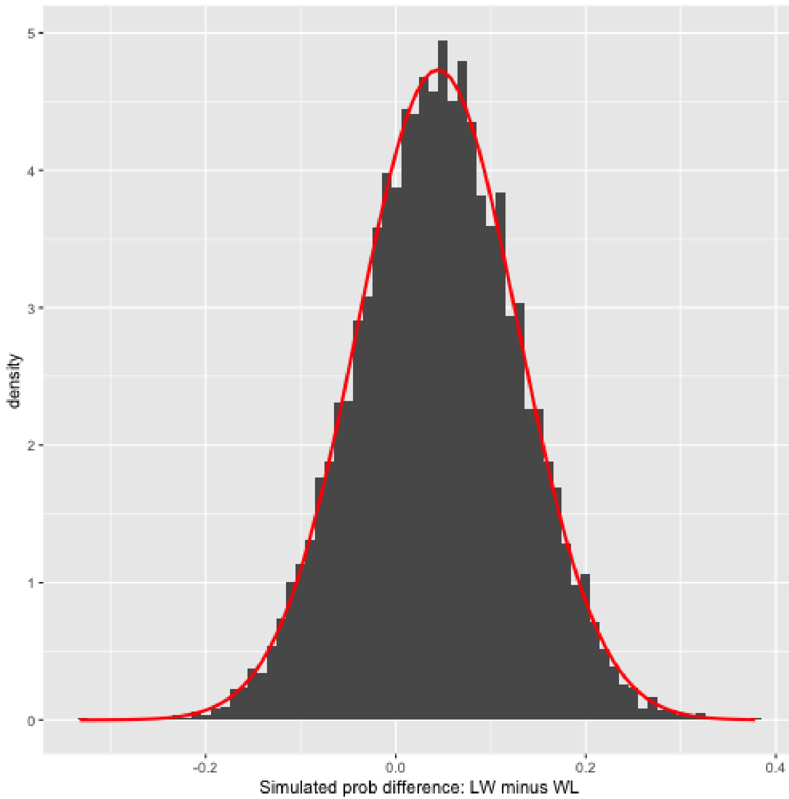
Update after losing game 1



Condition on a winning path



Estimate differences in means



Takeaways

- Game 2 looks like the most influential game, but we need more playoff data to say this with confidence!
- Bayesian thinking lends itself well to cases like playoffs where an uncertain outcome is determined by a series of events that happen over time
- It's very easy to update beta priors with Bernoulli trial outcomes
- Posteriors help visualize how uncertainty decreases as well as how mean estimate changes

References

- Understanding empirical Bayes estimation (using baseball statistics) by David Robinson
- ELO Rating System