Priors & The Power of Fake Data

Suppose the Dodgers have won W and lost L games thus far in the season.
How would you predict their end of season win percentage WP?

- 162 total games in the season
- no access to their schedule (e.g., ignore strength of schedule)
- Without using previous season's data (i.e. no Regression).

Guess their end of season win percentage.
Naive guess (ask any rando on the street):

$$\widehat{WP} = \frac{W}{W+L}$$

What's wrong with this?
When Dodgers have only played a few games, this estimate is bad.

Ex $W=3, L=0, \widehat{WP}=1$

Idea: Add fake data.

Suppose the Dodgers begin the season with W' wins and L' losses.

New guess:

$$\widehat{WP}' = \frac{W+W'}{W+W'+L+L'}$$

For concreteness:

$W=3,\ L=0,\quad \widehat{WP}=1$

Tom Tango: $W'=L'=15$ is good

$W=3,\ L=0,\ W'=15,\ L'=15,\quad \widehat{WP}'=\frac{18}{33}\approx .55$

quite different predictions early in the season

$W=45,\ L=30,\quad \widehat{WP}=\frac{45}{75}=.6$

$W=45,\ L=30,\ W'=15,\ L'=15,\quad \widehat{WP}'=\frac{60}{105}\approx .57$

similar predictions late in the season

Which is better?

Formalize this

Dodgers play $n = 162$ games in a season.

Suppose, for simplicity, that the Dodgers win each game with probability p.

Game outcomes $\{X_1, \ldots, X_n\}$, where

$$X_i \sim \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases} \stackrel{d}{=} \text{Bernoulli}(p)$$

Suppose we have observed m games thus far in the season.

Observed data $\{X_1, \ldots, X_m\}$. Each X_i is 1 or 0.

Observed # wins $W = \sum_{i=1}^{m} X_i$.

So, $W \sim \text{Binomial}(m, p)$

m = # trials (games)
p = prob. success (win)

and end-of-szn win percentage $WP \sim \frac{1}{n} \text{Binomial}(n, p)$

<u>Idea</u>. use observed data to estimate p, call it $\hat{p}$

Then, estimate $\widehat{WP} = \frac{1}{n} \mathbb{E}[\text{Binomial}(n, \hat{p})] = \frac{1}{n} \cdot n\hat{p} = \hat{p}$.

Maximum Likelihood estimate (MLE)

Choose $\hat{p}$ to be the value of p which maximizes the probability of observing the game outcomes $\{x_1, \ldots, x_m\}$ that we observed.

$$\hat{p}_{MLE} = \underset{p}{\operatorname{argmax}} \underbrace{\mathbb{P}(x_1, \ldots, x_m \mid p)}$$

likelihood : $\mathbb{P}$(data given parameter)

$$= \underset{p}{\operatorname{argmax}} \; \mathbb{P}(x_1 \mid p) \cdot \mathbb{P}(x_2 \mid p) \cdot \ldots \cdot \mathbb{P}(x_m \mid p)$$

by independence

$$= \underset{p}{\operatorname{argmax}} \prod_{i=1}^{m} \mathbb{P}(x_i \mid p)$$

by def of product

$$= \underset{p}{\operatorname{argmax}} \prod_{i=1}^{m} p^{x_i} (1-p)^{1-x_i}$$

because $x_i \sim Ber(p)$

$x_i = 1$ means $p^{x_i}(1-p)^{1-x_i} = p$

$x_i = 0$ means $p^{x_i}(1-p)^{1-x_i} = 1-p$

$$= \underset{p}{\operatorname{argmax}} \; p^{\sum_{i=1}^{m} x_i} (1-p)^{\sum_{i=1}^{m} (1-x_i)}$$

$$= \underset{p}{\operatorname{argmax}} \; p^W (1-p)^L$$

where $W = \sum_{i=1}^{m} x_i$ = number of wins (ones)

$L = \sum_{i=1}^{m} (1-x_i)$ = number of losses (zeros)

$$= \underset{p}{\operatorname{argmax}} \; \log \left[p^W \cdot (1-p)^L \right]$$

because log is monotonic increasing
to maximize $f(p)$ is to maximize $\log f(p)$

$$= \underset{p}{\operatorname{argmax}} \; W \log p + L \log (1-p)$$

to maximize the function $p \mapsto W \log p + L \log(1-p)$
take the derivative and set it equal to 0
(and check that the 2nd derivative is negative).

$$\frac{d}{dp} \left[W \log p + L \log(1-p) \right]$$

$$= W \cdot \frac{1}{p} - L \cdot \frac{1}{1-p} = 0$$

$$\Rightarrow \frac{W}{p} = \frac{L}{1-p} \Rightarrow p = \frac{W}{L}(1-p)$$

$$\Rightarrow p\left(1 + \frac{W}{L}\right) = \frac{W}{L} \Rightarrow p = \frac{\frac{W}{L}}{1+\frac{W}{L}}$$

$$\Rightarrow \hat{p}_{MLE} = \frac{W}{W+L}$$

same formula from earlier !!

The MLE is simply the observed win percentage midway through the season!

But we know this is a bad estimate early in the season.

So, why did the MLE go wrong??
How do we add the fake data W', L' to the MLE to get $\frac{W+W'}{W+W'+L+L'}$??

Before, to improve our estimate of WP, we added some fake data (W', L').

In adding fake data, we used prior information:
prior to the season, we assumed the Phillies have W' wins and L' losses.

What is a way of formalizing prior information?

Bayesian statistics – the belief/philosophy that we should treat a parameter (e.g. p) as having a probability distribution

Frequentist statistics – treats a parameter as an unknown fixed number

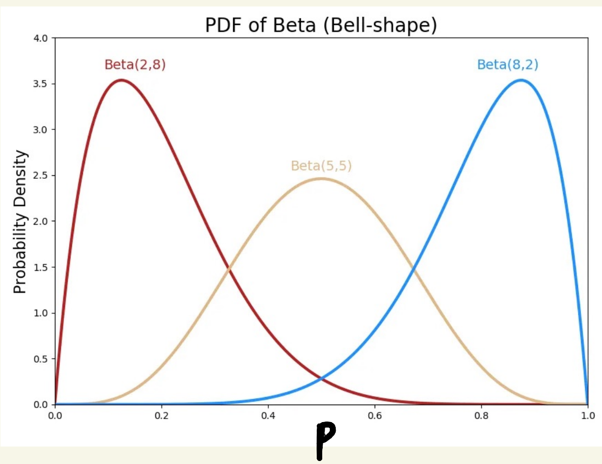
So, our way of formalizing the addition of prior "fake" data is to, prior to seeing the data, give a probability distribution to the parameter (e.g. p) which reflects our prior belief on what p is more likely to be than not!

Formally, we use the Beta-Binomial model:

$$\begin{cases} W \sim \text{Binomial}(m, p) \\ p \sim \text{Beta}(\alpha, \beta) \rightarrow \text{Prior} \\ \qquad \alpha = W' + 1, \quad \beta = L' + 1 \end{cases}$$

Beta distribution has density $f(p \mid \alpha, \beta) = C \cdot p^{\alpha-1}(1-p)^{\beta-1}$

on the interval $p \in [0,1]$, where C is a constant chosen so that the distribution integrates to 1.



For example, $p \sim \text{Beta}(5,5)$ encodes a preference that p is closer to 0.5

As before, we wish to estimate p, this time with a
Maximum a-Posteriori (MAP) Estimate:

Choose the $\hat{p}$ which maximizes the Posterior PRobability of p.

Bayesian Approach to Parameter Estimation
1. PRioR
2. observe data
3. adjust our posterior dist for p given the data

$$\hat{p}_{MAP} = \underset{p}{\operatorname{argmax}} \underbrace{\mathbb{P}(p|W)}_{\text{Posterior} = \mathbb{P}(\text{parameter} | \text{data})}$$

$$= \underset{p}{\operatorname{argmax}} \frac{\mathbb{P}(W|p) \cdot \mathbb{P}(p)}{\mathbb{P}(W)} \quad \text{by Bayes' Rule}$$

$$= \underset{p}{\operatorname{argmax}} \underbrace{\mathbb{P}(W|p)}_{\text{likelihood}} \cdot \underbrace{\mathbb{P}(p)}_{\text{prior}}$$

since $\mathbb{P}(W)$ has no p term

$$= \underset{p}{\operatorname{argmax}} \mathbb{P}(\text{Binomial}(n,p) = W) \cdot \mathbb{P}(\text{beta}(\alpha,\beta) = p)$$

$$= \underset{p}{\operatorname{argmax}} \binom{m}{W} p^{W} (1-p)^{m-W} \cdot C\, p^{\alpha-1} (1-p)^{\beta-1}$$

$$= \underset{p}{\operatorname{argmax}}\; p^{W} (1-p)^{L} \cdot p^{\alpha-1} (1-p)^{\beta-1}$$

$$= \underset{p}{\operatorname{argmax}}\; p^{W+\alpha-1} (1-p)^{L+\beta-1}$$

$=$ ••• same process as before

$$= \frac{W+\alpha-1}{W+\alpha-1+L+\beta-1}$$

$$= \frac{W+W'}{W+W'+L+L'} \quad \text{if} \quad \begin{array}{l} W' = \alpha-1 \\ L' = \beta-1 \end{array}$$

The MAP estimate is simply the win percentage if we add $\alpha-1$ fake wins and $\beta-1$ fake losses!!

Can use past seasons to tune a smart choice for α, β.

Note: $\alpha=1, \beta=1 \Rightarrow \hat{p}_{MAP} = \hat{p}_{MLE}$

add no fake data

Model

$$\begin{cases} W \sim \text{Binomial}(n, p) \\ p \sim \text{Uniform}[0,1] \end{cases}$$

$\rightarrow$ uninformative PRIOR which encodes no preference on p

$$\hat{p}^{(MAP)} = \underset{p}{\operatorname{argmax}}\ \mathbb{P}(p|W) = \underset{p}{\operatorname{argmax}}\ \mathbb{P}(W|p) \cdot \underbrace{\mathbb{P}(p)}_{1}$$

$$= \underset{p}{\operatorname{argmax}}\ \mathbb{P}(W|p) = \hat{p}^{(MLE)} = \frac{W}{W+L}$$

Takeaways

- Bayesian Statistics: treat a parameter (e.g., p) as having a distribution
- Blend observed data with PRIOR knowledge, encoding info not seen in the data, to make better predictions