The Power of Fake Data (Priors)

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

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

Problem?
- injuries
- no strength of schedule

$\Big\}\rightarrow$ ignore

- lack of data $\longrightarrow$ $W=3, L=0, \widehat{WP}=1$

Idea Add Fake Data!

$$\widehat{WP}' = \frac{W+W'}{W+W'+L+L'} \qquad \text{fake } (W', L')$$

Tom Tango: $W' = 15,\ L' = 15$

$W=3, L=0, \widehat{WP}' = \frac{18}{33} \approx .55,\ \widehat{WP} = 1$

Some sort of Regression happening here

Shrinking obs. WP $\frac{3}{3+0}$ to 50% $\frac{15}{15+15}$

Can we formalize this?

Phillies play $n=162$ games

Simple model: Phillies win each game w.p. p

Game outcomes $X_1, \dots, X_n$

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

We've observed just m of these game outcomes $X_1, \dots, X_m$.

Observed # wins $\quad W = \sum_{i=1}^{m} X_i \sim \text{Binomial}(m, p)$.

* At the same time, the actual end-of-szn WP will be

$$\frac{1}{n}\text{Binomial}(n, p) = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

has expected value p.

So our task is to estimate p from our observed games.

* We will discuss a few ways to estimate p.

Maximum Likelihood Estimate (MLE)

Choose the $\hat{p}$ which maximizes the probability of observing the data that we observed.

$$\hat{p}^{(MLE)} = \underset{p}{\operatorname{argmax}} \; \underbrace{\mathbb{P}(X_1, \ldots, X_m \mid p)}_{\text{each } X_i \text{ Bernoulli}(p)}$$

$$= \underset{p}{\operatorname{argmax}} \; \mathbb{P}(X_1 \mid p) \cdot \mathbb{P}(X_2 \mid p) \cdot \ldots \mathbb{P}(X_m \mid p)$$

by independence

$$= \underset{p}{\operatorname{argmax}} \; \underbrace{\prod_{i=1}^{m}}_{\text{product}} \mathbb{P}(X_i \mid p)$$

$$\mathbb{P}(X_i \mid p) \rightarrow \begin{cases} \mathbb{P}(X_i = 0 \mid p) = 1 - p \\ \mathbb{P}(X_i = 1 \mid p) = p \end{cases}$$

$$\mathbb{P}(X_i \mid p) = p^{X_i}(1-p)^{1-X_i}$$

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

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

$$\prod_{i=1}^{m} p^{x_i}(1-p)^{1-x_i} = p^{x_1} p^{x_2} \cdots p^{x_m} (1-p)^{1-x_1}(1-p)^{1-x_2} \cdots (1-p)^{1-x_m}$$

$$= p^{x_1+x_2+\cdots+x_m} (1-p)^{(1-x_1)+\cdots+(1-x_m)}$$

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

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

$\log x$ – monotonic increasing

$$= \underset{p}{\arg\max} \quad W \cdot \log p + L \cdot \log(1-p)$$

How do we maximize this?

set derivative equal to 0 and solve.

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

$$\Longrightarrow \quad W \frac{1}{p} = L \frac{1}{1-p}$$

$$W(1-p) = Lp$$

$$W = p(W+L)$$

$$p = \frac{W}{W+L}$$

$$\boxed{\hat{p}^{(MLE)} = \underset{p}{\arg\max} \; \mathbb{P}(X_1, \ldots, X_m \mid p) = \frac{W}{W+L}}$$

We know this is bad.

Why did the MLE go wrong?

Vaguely though we know we are looking to add fake data (W', L') to acheive

$$\frac{W+W'}{W+W'+L+L'}.$$

In adding fake data, we use 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 a parameter (e.g. p) itself has an (unknown) distribution

Frequentist Statistics — treats a parameter as a fixed (unknown) number

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

$p \sim$ PRIOR Distribution

Our way of formalizing the addition of prior "fake data" is to, prior to seeing any 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.

Tango: added $W' = 15$, $L' = 15 \rightarrow p \approx \frac{15}{30} = \frac{1}{2}$

$$\widehat{wp} = \frac{W + 15}{W + L + 30}$$

early in szn closer to 1/2

later in szn closer to obs. wp

PRIOR: loosely, let's be closer to $\frac{1}{2}$ and further from the extremes 0 and 1

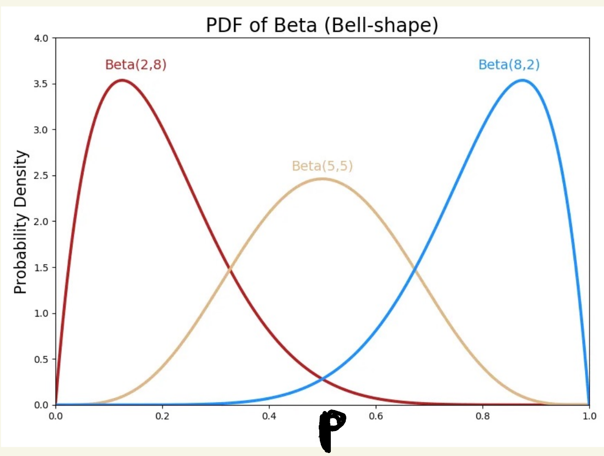
What prior dist. should we use for p?

$p \in [0,1]$

Simplest dist. defined on $[0,1]$ which has flexible parameters? Beta

$$\begin{cases} W \sim \text{Binomial}(n, p) \\ p \sim \text{Beta}(\alpha, \beta) \longrightarrow \text{Prior Distribution} \end{cases}$$

Beta-Binomial model



y-axis: $\mathbb{P}(\text{Beta}(\alpha,\beta) = p)$

$p \sim \text{Beta}(\alpha, \beta)$

$p \in [0,1]$

Beta-dist. has density

$$\mathbb{P}(\text{Beta}(\alpha,\beta) \in [p, p+dp)) = f(p \mid \alpha, \beta)$$

$$= C \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}$$

where C is a constant so that $\int_0^1 f(p)\,dp = 1$.

Have model $\begin{cases} W \sim \text{Binomial}(n,p) \\ P \sim \text{Beta}(\alpha,\beta) \longrightarrow \text{Prior Distribution} \end{cases}$

Want to estimate p.

Before we ever introduced a PRIOR:

Maximum Likelihood Estimate (MLE)

Choose the $\hat{p}$ which maximizes the probability of observing the data that we observed.

$$\hat{p}^{(MLE)} = \underset{p}{\text{argmax}}\ \mathbb{P}(X_1, \dots, X_m \mid p)$$

Now that have a prior:

Maximum a-Posteriori (MAP)

Choose the $\hat{p}$ which maximizes the posterior probability of p (the probability of p using the information from our observed data).

Bayesian Approach to parameter estimation

1. prior $p \sim \text{Beta}(\alpha, \beta)$
2. observe data $X_1, \dots, X_m$
3. adjust our dist. for p using the data $\mathbb{P}(p \mid X_1, \dots, X_m)$

$$\hat{p}^{(MAP)} = \underset{p}{\text{argmax}} \; \underbrace{\mathbb{P}(p \mid W, L)}_{\text{the posterior prob. of a parameter is the prob. of that parameter given the observed data}}$$

Bayes Rule

$$= \underset{p}{\operatorname{argmax}} \frac{\mathbb{P}(W, L \mid p) \cdot \mathbb{P}(p)}{\mathbb{P}(W, L)}$$

$$= \underset{p}{\operatorname{argmax}} \underbrace{\mathbb{P}(W, L \mid p)}_{\text{Binomial}} \cdot \underbrace{\mathbb{P}(p)}_{\text{Beta}}$$

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

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

$$\mathbb{P}(W, L \mid p) = \mathbb{P}\left(\text{obs. \# wins} = W, \text{ obs. \# losses} = L \mid p\right)$$

$$= \mathbb{P}\left(\sum_{i=1}^{m} X_i = W, \; m - \sum_{i=1}^{m} X_i = L \mid p\right)$$

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

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

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

$$\left\{ W \sim \text{Binomial}(m,p) \right. \Longrightarrow \hat{p}^{(MLE)} = \frac{W}{W+L}$$

$$\left\{ \begin{array}{l} W \sim \text{Binomial}(m,p) \\ p \sim \text{Beta}(\alpha,\beta) \\ \quad \alpha = W'+1, \quad \beta = L'+1 \end{array} \right. \Longrightarrow \hat{p}^{(MAP)} = \frac{W+W'}{W+W'+L+L'}$$

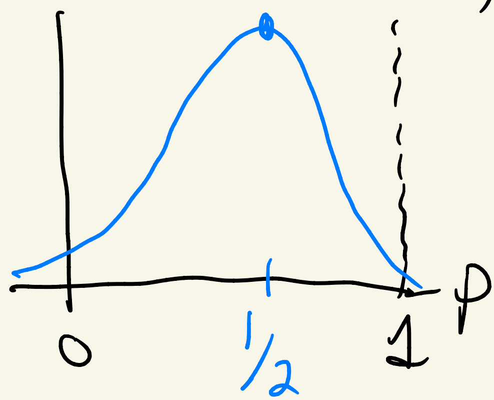
$$\left\{ \begin{array}{l} W \sim \text{Binomial}(m,p) \\ p \sim \text{Uniform}[0,1] \end{array} \right. \Longrightarrow \hat{p}^{(MAP)} = \hat{p}^{(MLE)} = \frac{W}{W+L}$$

Takeaways

- Bayesian approach: treat parameter (e.g. p) as having a distribution

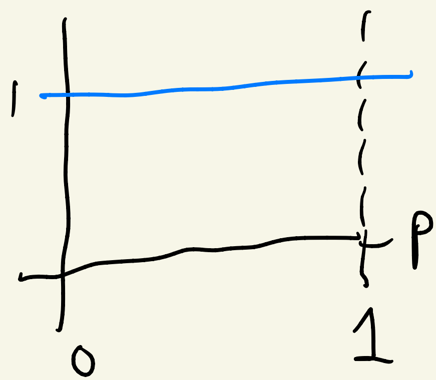
- PRIOR: allows us to make better predictions by encoding information not seen in the available data

* If you have prior reason to believe, before looking at data, that p should be $\approx \frac{1}{2}$ when small # of obs., then choose a prior like Beta(16, 16) which looks like



* Conversely, if you have <u>no prior info. or beliefs</u> on what p should be,

Uniform[0, 1]



<u>model</u> $\begin{cases} W \sim \text{Binomial}(m, p) \\ p \sim \text{Uniform}[0,1] \end{cases}$ prior

$$\hat{p}^{(MAP)} = \underset{p}{\operatorname{argmax}} \ \mathbb{P}(p \mid W)$$

$$= \underset{p}{\operatorname{argmax}} \ \mathbb{P}(W \mid p) \cdot \underbrace{\mathbb{P}(p)}$$

$$\mathbb{P}(p) = \mathbb{P}(\text{Uniform}[0,1] = p) = 1$$

$$= \underset{p}{\operatorname{argmax}} \ \mathbb{P}(W \mid p)$$

$$= \hat{p}^{(MLE)} = \frac{W}{W+L}.$$