Lec 19 MATH 341/641

Little break from Bayesian ~~to~~ discuss a nice example of the BetaBinomial fit. Consider the following dataset. 6,115 sets of parents. Each couple had ≥ 12 children. We record the gender of the 12 children by couple. Below is # of boys:

# Boys	0	1	2	3	4	5	6	7	8	9	10	11	12	Total
X	3	24	104	286	670	1033	1343	1112	829	478	181	45	7	6115
Binomial Pred.	1	12	72	258	628	1085	1367	1266	854	410	133	26	2	
BetaBinom Pred														

To fit a binomial model assume DGP: $X_1, \ldots, X_{73380} \overset{iid}{\sim} \text{Bern}(\theta)$

and use $\hat{\theta}^{MLE} = \bar{x} = \dfrac{38100}{73800} = .519$, expected as gender ratio is always tilted a bit towards males and nobody knows why

We calculate

$AIC = 101620$

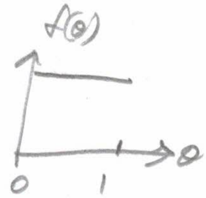
To fit a betabinomial model, assume DGP: $X_1, \ldots X_{6115} \overset{iid}{\sim} \text{BetaBinom}(n=12, \alpha, \beta)$

We fit $\hat{\alpha}^{MLE} = 34.06$, $\hat{\beta}^{MLE} = 31.54$ and calculate

$AIC = 27990 \implies$ Much Better Fit!!!!

What does this betabinomial model truly mean??

Priors are subjective. But can we make them "objective" to minimize their effect on the inference. Have cake and eat it too.



Laplace's prior is "flat" i.e. "it doesn't give any special preference to any of the θ's"

But we also say it implies $E[\theta] = 0.5$ and $n_0 = 2$ one $x=1$, one $x=0$.

$$f(\theta \mid x) = \text{Beta}(\overset{1}{\alpha} + x,\ \overset{1}{\beta} + n - x)$$

This doesn't seem too objective! Further...

Let's say we care about inference for $\phi = g(\theta) = \frac{1-\theta}{\theta}$, odds against

1:1 transformation

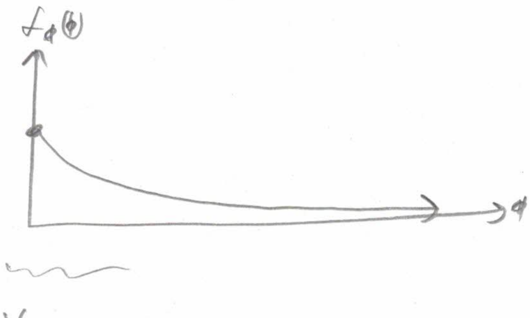
What does Laplace's prior mean on this new scale?

$$f_\phi(\phi) = f_\theta(g^{-1}(\phi)) \left| \frac{d}{d\phi}\left[g^{-1}(\phi)\right] \right| = \mathbb{1}_{\frac{1}{\phi+1} \in (0,1)} \frac{1}{(\phi+1)^2} = \frac{1}{(\phi+1)^2} \mathbb{1}_{\phi+1 \in (1,\infty)} = \frac{1}{(\phi+1)^2} \mathbb{1}_{\phi \in (0,\infty)}$$

$$\phi = \frac{1}{\theta} - 1 \Rightarrow \phi + 1 = \frac{1}{\theta} \Rightarrow \theta = \frac{1}{\phi+1} = g^{-1}(\phi) \quad \left| \frac{d}{d\phi}\left[-(\phi+1)^{-2}\right] \right| = \frac{1}{(\phi+1)^2}$$

$\propto F_{2,2}$ dist

$= \text{BetaPrime}(1,1)$



which is not flat!! Further, there is no way to have a flat PDF on $(0,\infty)$!

You are saying small odds against values are more likely than large

Laplace indifference prior is tied to a specific parametrization

What happens in this case??

FYI... what is posterior?

$$f(\phi|x) \propto p(x|\phi)\, f(\phi) \propto \binom{n}{x}\left(\frac{1}{\phi+1}\right)^{x}\left(\frac{\phi}{\phi+1}\right)^{n-x}\frac{1}{(\phi+1)^{2}}\mathbb{1}_{\phi\in(0,\infty)} \propto \frac{\phi^{n-x}}{(\phi+1)^{n+2}}\mathbb{1}_{\phi\in(0,\infty)}$$

$$\propto \text{BetaPrime}(n-x+1,\, x+2)$$

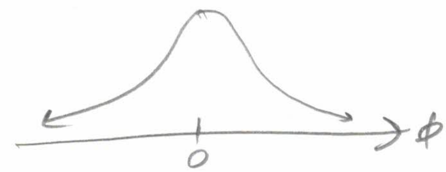
BetaPrime is the conjugate prior for the Binomial PDF parameterized with odds-against log-odds-against

What about $\phi = \ln\left(\frac{1-\theta}{\theta}\right) = \underbrace{\ln\left(\frac{1}{\theta}-1\right)}_{=g(\theta)} \Rightarrow e^\phi = \frac{1}{\theta}-1 \Rightarrow e^\phi+1 = \frac{1}{\theta} \Rightarrow \theta = \frac{1}{e^\phi+1}$ $= g^{-1}(\phi)$

$$\left|\frac{d}{d\phi}\left[g^{-1}(\phi)\right]\right| = \frac{e^\phi}{(e^\phi+1)^2}$$

$$f_\phi(\phi) = f_\theta\left(g^{-1}(\phi)\right)\frac{d}{d\phi}\left[g^{-1}(\phi)\right]$$

$$= \mathbb{1}_{\underbrace{\underbrace{\underbrace{\frac{1}{e^\phi+1}\in(0,1)}_{e^\phi+1\in(1,\infty)}}_{e^\phi\in(0,\infty)}}_{\phi\in\mathbb{R}}} \frac{e^\phi}{(e^\phi+1)^2} = \text{Logistic}(0,1)$$



Not flat either!!!

Posterior?

$$f(\phi|x) \propto p(x|\phi)f(\phi) = \binom{n}{x}\left(\frac{1}{e^\phi+1}\right)^x\left(\frac{e^\phi}{e^\phi+1}\right)^{n-x}\frac{e^\phi}{(e^\phi+1)^2} \propto \frac{(e^\phi)^{n-x+1}}{(e^\phi+1)^{n+2}} \propto \text{Type4Logistic}(n-x+1,\ x+1)$$

Type 4 Logistic is the conjugate prior for the Binomial DGP parameterized by log odds.

Is there a procedure that gives us the following? Assume DGP $\Rightarrow p(\vec{X}|\theta)$. For any 1:1 transformation g on θ, can we have...

$$\begin{array}{ccc} p(\vec{X}|\theta) & \xrightarrow{\text{procedure}} & f(\theta) \\ g\downarrow\uparrow g^{-1} & & g\downarrow\uparrow g^{-1} \\ p(\vec{X}|\phi) & \xrightarrow{\text{procedure}} & f(\phi) \end{array}$$

This means you have the same prior regardless of whatever $\phi = g(\theta)$ you employ. This procedure ~~was~~ discovered by Jeffreys in the 1930's.

Thm: $p_J(\theta) \propto \sqrt{I_1(\theta)}$ satisfies the spec, above.

$$\begin{array}{ccc} p(\vec{X}|\theta) & \xrightarrow[\text{Procedure}]{\text{Jeffreys}} & f_J(\theta) \propto \sqrt{I_1(\theta)} \\ g\downarrow\uparrow g^{-1} & & g\downarrow\uparrow g^{-1} \\ p(\vec{X}|\phi) & \xrightarrow[\text{Procedure}]{\text{Jeffreys}} & f_J(\phi) \propto \sqrt{I_1(\phi)} \end{array}$$

true by change of variables

$$f_J(\theta) \propto \sqrt{I_1(\theta)} \Longleftrightarrow f_J(\phi) \propto \sqrt{I_1(\phi)}$$

Needs to be proven

Proof that $f_J(\theta) \propto \sqrt{I_n(\theta)} \Rightarrow f_J(\phi) \propto \sqrt{I_n(\phi)}$

$$f_J(\phi) = f_J\left(\overset{\theta}{\overset{\shortparallel}{g^{-1}(\phi)}}\right) \left| \frac{d}{d\phi}\left[\overset{\theta}{\overset{\shortparallel}{g^{-1}(\phi)}}\right] \right|$$

$$\propto \sqrt{I_n(\theta)} \left| \frac{d\theta}{d\phi} \right|$$

$$= \sqrt{I_n(\theta)\left(\frac{d\theta}{d\phi}\right)^2}$$

$$= \sqrt{E_X\left[\ell'(\theta; \vec{x})^2\right] \frac{d\theta}{d\phi}\frac{d\theta}{d\phi}}$$

$$= \sqrt{E_X\left[\frac{d\ell}{d\theta}\frac{d\ell}{d\theta}\frac{d\theta}{d\phi}\frac{d\theta}{d\phi}\right]}$$

$$= \sqrt{E_X\left[\frac{d\ell}{d\phi}\frac{d\ell}{d\phi}\right]}$$

$$= \sqrt{E_X\left[(\ell'(\phi; \vec{x}))^2\right]}$$

$$= \sqrt{I_n(\phi)}$$

The other direction's proof is the same (HW).

What is Jeffrey's prior for DGP: $X \sim \text{Bin}(n, \theta)$?

$$\mathcal{L}(\theta; x) = P(X|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

$$\ell(\theta; x) = \ln\left(\binom{n}{x}\right) + x\ln(\theta) + (n-x)\ln(1-\theta)$$

$$\ell'(\theta; x) = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

$$\ell''(\theta; x) = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}$$

$$I_n(\theta) = E[-\ell''(\theta; X)] = E\left[\frac{X}{\theta^2} + \frac{n-X}{(1-\theta)^2}\right] = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = n\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) = n\frac{1}{\theta(1-\theta)}$$

$$f_J(\theta) \propto \sqrt{I_n(\theta)} = \sqrt{n\frac{1}{\theta(1-\theta)}} \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} = \theta^{\frac{1}{2}-1}(1-\theta)^{\frac{1}{2}-1} \propto \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$$

Jeffrey's prior is different than Laplace's prior! AKA arcsine distr.

DGP: $X \sim \text{Bin}(n, \theta)$, $f(\theta) = \text{Beta}(\alpha, \beta) \Rightarrow f(\theta|x) = \text{Beta}(\alpha + x, \beta + n - x)$

α: # pseudosuccesses, β: # pseudofailures, $n_0 = \alpha + \beta$: # pseudotrials

Complete objectivity means $\alpha = \beta = n_0 = 0$ (Haldane's prior of Ignorance)

$\Rightarrow f(\theta) = \text{Beta}(0,0)$ Improper!

$\Rightarrow f(\theta|x) = \text{Beta}(x, n-x)$ proper if $x \geq 1$ and $n - x \geq 1$

Three objective conjugate priors: (AKA uninformative priors)

① Laplace $f(\theta) = \text{Beta}(1,1)$

② Jeffreys $f(\theta) = \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$

③ Haldane $f(\theta) = \text{Beta}(0,0)$

All three of these are commonly in use. If x, n are large... almost no difference in inference

Has this all been a hack to get what we want???

What about subjective / informative priors? When n_0 is large. Do these have any usefulness? Yes! Here is my favorite example.

DGP: $X \sim \text{Bin}(n, \theta)$, $f(\theta) = \text{Beta}(\alpha, \beta)$

n: # at bats
X: # of hits
θ: lifetime career batting avg

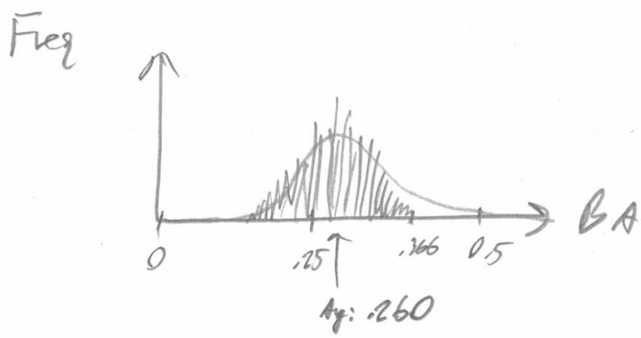
Consider a new hitter. $n = 3$, $x = 2 \Rightarrow \hat{\theta}^{MLE} = \frac{2}{3} = .667$

This is a terrible estimate! Nobody bats .667. Highest Batting Average in history is Ty Cobb $\theta = .366$

The three objective priors don't help you! What to do?

Leverage historical data!

Let's examine all BA's where $n \geq 500$. Histogram:



Now fit a beta distr to it since we want to have a conjugate prior

$\hat{\alpha}^{MLE} = 78.7$, $\hat{\beta}^{MLE} = 224.8$

Let $f(\theta) = \text{Beta}(78.7, 224.8)$

$E(\theta) = .260$, $n_0 = 303.5$

This prior has the weight of 303.5 at bats

Now lets estimate!

$$f(\theta|x) = \text{Beta}(\alpha+x, \beta+n-x) = \text{Beta}(78.7+x, 224.8+n-x)$$

For our batter, $= \text{Beta}(80.7, 225.8)$

$$\hat{\theta}^{MMSE} = \frac{80.7}{80.7+225.8} = .263$$ good estimate!!

Shrinkage? $\rho = \frac{\alpha+\beta}{\alpha+\beta+n} = \frac{n_0}{n_0+n} = \frac{3035}{3035+2} = 99\%$ tons of shrinkage which is good!

When can you use informative priors?

1. when n is small
2. when you are confident that this estimation does not differ from historical patterns.

This is a major advantage of the Bayesian system!

DONE with Beta-Binomial Model!