Lecture 10 3/14/17 Math 341 $I(\theta) := \text{Var}_X[S(\theta;x)] = \ldots = E_X[S(\theta;x)^2] = \ldots = E[-\ell''(\theta;x)]$ 1

$I(\theta)$ measures how much information is in X for a r.v.

Let's see this for $X \sim \text{Binom}(n,\theta)$ for fixed n

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

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

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

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

recall $E[aX+c] = aE(X)+c$

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

$$= n\left(\frac{1}{\theta(1-\theta)}\right)$$

Not a function of X! X is averaged over...

If $\theta = \frac{1}{2}$, $n=1$, How much info? $I(\frac{1}{2}) = 4$ ← the r.v. does not have too much info about θ on average

If $\theta = \frac{1}{100}$, $n=1$ ....

$I\left(\frac{1}{100}\right) = 101.01$ ← the r.v. has a ton of info

Why?

Why should there be a mult. function of n? More data $\Rightarrow$ more info. Remember binomial is n bernoullis... more bernoulli draws $\Rightarrow$ more info

---

Back to the issue... CONSIDER:

What if $p(\theta) \propto \sqrt{I(\theta)}$ AKA the Jeffrey's prior

For $p(X|\theta) = \mathrm{Bin}(n,\theta) \Rightarrow I(\theta) = n \frac{1}{\theta(1-\theta)}$

$$\Rightarrow p(\theta) \propto \sqrt{n \frac{1}{\theta(1-\theta)}} \propto \underbrace{\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}}_{\text{Kernel of the Beta density!}} \propto \mathrm{Beta}\left(\tfrac{1}{2},\tfrac{1}{2}\right) = \frac{1}{B\left(\frac{1}{2},\frac{1}{2}\right)} \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$$

(E.C.) $\pi = B\left(\frac{1}{2},\frac{1}{2}\right)$

$$= \frac{1}{\pi\sqrt{\theta(1-\theta)}}$$

"Arcsine distr."

Jeffreys prior is an uninformative ($\alpha+\beta=1$ i.e. small)

Does it do its job? Lets reparametrize to odds...

scalar
$p(\theta|x) \rightarrow p(\theta)$
$\phi = t(\theta)$
scalar
$p(\phi|x) \rightarrow p(\phi)$
does
$p_\phi(\phi) = p_\theta(t^{-1}(\phi)) \left|\frac{d}{d\phi}(t^{-1}(\phi))\right|$
?

let $R = t(\theta) = \frac{\theta}{1-\theta}$ and $\theta = t^{-1}(R) = \frac{R}{R+1}$

$$p(X|R) = \binom{n}{x}\left(\frac{R}{R+1}\right)^x \underbrace{\left(1-\frac{R}{R+1}\right)}_{\frac{1}{R+1}}^{n-x} = \binom{n}{x}\frac{R^x}{(R+1)^n}$$

$p(R) = ???$

$$\ell(R;X) = \ln\binom{n}{x} + x\ln(R) - n\ln(R+1)$$
$$\ell'(R;X) = \frac{x}{R} - \frac{n}{R+1}$$
$$\ell''(R;X) = -\frac{x}{R^2} + \frac{n}{(R+1)^2}$$

$I(R) =$
$$E\left(-\ell''(R;X)\right) = \frac{E(X)}{R^2} - \frac{n}{(R+1)^2} = \frac{n\frac{\cancel{R}}{R+1}}{R^{\cancel{2}}} - \frac{n}{(R+1)^2} = n\left(\frac{1}{R(R+1)} - \frac{1}{(R+1)^2}\right) = n\frac{1}{R(R+1)^2}$$

$$p(R) \propto \sqrt{n\frac{1}{R(R+1)^2}} = \underbrace{\frac{1}{\sqrt{R}}\frac{1}{R+1}}_{g(R)} \qquad \int_0^\infty g(R)\,dR = \pi \Rightarrow p(R) = \frac{1}{\pi}\frac{1}{\sqrt{R}}\frac{1}{R+1}$$

Now use change of var's!

//

$$p(R) = p_\theta(t^{-1}(R))\left|\frac{d}{dR}\left[t^{-1}(R)\right]\right| = \frac{1}{\pi}\left(\frac{R}{R+1}\right)^{-\frac{1}{2}}\underbrace{\left(1-\frac{R}{R+1}\right)}_{\frac{1}{R+1}}^{-\frac{1}{2}}\left|\frac{1}{(R+1)^2}\right| = \frac{1}{\pi}\frac{1}{\sqrt{R}}\cancel{(R+1)}\frac{1}{(R+1)^{\cancel{2}}}$$

How is this possible? Given...

$$p(X|\theta),\ p(X|\phi),\quad \phi = t(\theta),\ \theta = t^{-1}(\phi)$$

Under Jeffreys strategy,

$$p(\theta) \propto \sqrt{I(\theta)} \quad \text{and} \quad p(\phi) \propto \sqrt{I(\phi)}$$

Show that:

$$p(\phi) = p_\theta\Big(\overset{\theta}{\overset{\shortparallel}{t^{-1}(\phi)}}\Big) \left| \frac{d}{d\phi}\Big[\overset{\theta}{\overset{\shortparallel}{t^{-1}(\phi)}}\Big] \right| \propto \sqrt{I(\phi)}$$

$$= p_\theta(\theta) \left| \frac{d\theta}{d\phi} \right|$$

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

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

$$= \sqrt{E\left[S(\theta; X)^2\right] \frac{d\theta}{d\phi} \cdot \frac{d\theta}{d\phi}}$$

$$= \sqrt{E\left[\frac{d\ell}{\cancel{d\theta}} \cdot \frac{d\ell}{\cancel{d\theta}} \cdot \frac{\cancel{d\theta}}{d\phi} \cdot \frac{\cancel{d\theta}}{d\phi} \right]}$$

$$= \sqrt{E\left[\left(\frac{d\ell}{d\phi}\right)^2\right]}$$

$$= \sqrt{E\left[S(\phi; x)^2\right]}$$

$$= \sqrt{I(\phi)} \quad \checkmark$$

$\theta \sim \text{Beta}(0,0)$ Haldane

$\theta \sim \text{Beta}(\frac{1}{2},\frac{1}{2})$ Jeffreys

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

all uninformative

+'s, -'s to each...

up to you...

Jeffreys used Fisher's own invention against him!!

New concept:

You are trying to guess a baseball player's the B.A., θ. The simple BA is

$$\hat{\theta} = BA := \frac{\#\text{HITS}}{\#\text{AT BATS}} = \frac{x}{n}$$

with some approx's ... we use the model ..

$$\#\text{HITS} \approx \text{Bin}(\#\text{at bats}, \theta)$$

BA is the $\hat{\theta}_{MLE}$

When does $\hat{\theta}_{MLE}$ have poor performance? If n is small. lets say $n=2$

$n=2,\ x=0 \Rightarrow \hat{\theta}=0,\ x=1 \Rightarrow \hat{\theta}=0.5,\ x=2 \Rightarrow \hat{\theta}=1$ all absurd!

Solution! Shrink! Use $\theta \sim \text{Beta}(\alpha, \beta)$

$\hat{\theta}_{MMSE} = \frac{x+\alpha}{\alpha+\beta+n}$ which includes a shrinkage towards $\frac{\alpha}{\alpha+\beta}$ with weight $\frac{\alpha+\beta}{n+\alpha+\beta}$

How to pick prior?

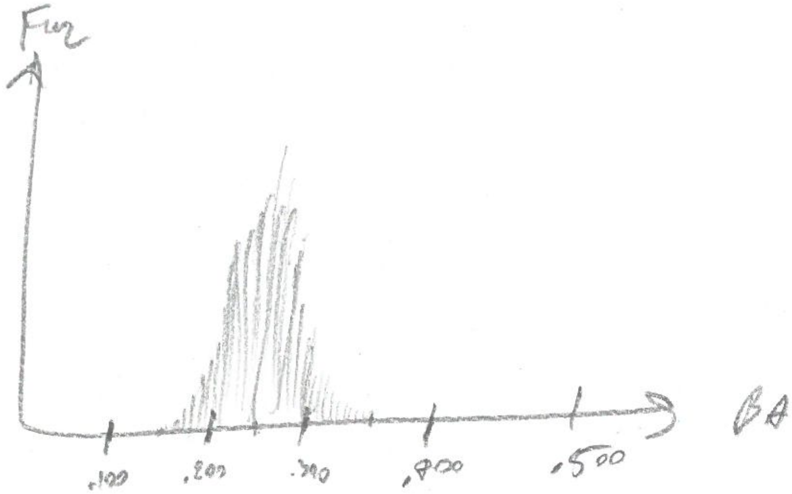
$\theta \sim \text{Beta}(1,1)$ Shrink towards 0.5 absurd!

How about look at all historical BA's for tons of players!

i.e. use prior data to build a prior $\Rightarrow$ empirical Bayes.

Here's how it works. Get my reasonable $\hat{\theta}$'s for previous players

lets say $n > 500$ at bats.



Fit a beta to the prior data. Using MLE's, $\hat{\alpha}_{MLE} = 78.7$, $\hat{\beta}_{MLE} = 224.8$

This has the strength of $n = 303.5$ at bats $\Rightarrow$ Strong!

$$\hat{\theta}_{MMSE} = \frac{x + 78.7}{n + 303.5}$$ which will perform better than $\hat{\theta}_{MLE} = \frac{x}{n}$ for n small.

Steps

1. Get old data
2. Fit conjugate prior data to it using MLE's
3. Use the fit hyperparams for inference

Done with beta, binomial ... on to new model..

$$\mathcal{F} = \text{Geometric} \Rightarrow X|\theta \sim \text{Geom}(\theta) := (1-\theta)^x \theta$$

x represents # failures

We see n trials .. $X = \{X_1, \ldots, X_n\}$

$$P(X;\theta) = \prod_{i=1}^{n} (1-\theta)^x \theta = (1-\theta)^{\sum x_i} \theta^n$$

~~$\hat{\theta}_{MLE} = \frac{1}{1+\bar{x}}$ we did this before~~

$\text{Supp}(X) = \{0, 1, \ldots\} = \mathbb{N}_0$

Param Space $\Theta = (0,1)$

$$E[X] = \sum_{x=0}^{\infty} x(1-\theta)^x \theta = \ldots = \frac{1}{\theta} - 1$$

Math 241 exercise

$\theta \uparrow \Rightarrow x \downarrow$, $\theta \downarrow \Rightarrow x \uparrow$

What is the Kernel of the posterior?

$$p(\theta|x) \propto p(x|\theta)p(\theta) = \underbrace{(1-\theta)^{\Sigma x_i}\,\theta^{n}}_{\text{Kernel of beta}} \quad p(\theta)$$

Now to find the conjugate prior... what to do? Pattern match!!

Let $p(\theta)$ be same kernel as $f(x|\theta)$

$\Rightarrow \theta \sim \text{Beta}(\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$\Rightarrow p(\theta|x) \propto \left((1-\theta)^{\Sigma x_i}\theta^{n}\right)\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right) = \theta^{n+\alpha-1}(1-\theta)^{\Sigma x_i+\beta-1}$$

$$\propto \text{Beta}(n+\alpha, \Sigma x_i+\beta)$$

$\Rightarrow$ Beta is also conjugate for the geometric likelihood model!

$\hat{\theta}_{MMSE} = \dfrac{n+\alpha}{n+\alpha+\Sigma x_i+\beta}$, $\hat{\theta}_{MAE} = q\text{beta}(.5, n+\alpha, \Sigma x_i+\beta)$, $\hat{\theta}_{MAP} = \dfrac{n+\alpha-1}{n+\alpha+\Sigma x_i+\beta-2}$ when appropriate

Do α,β have special pseudocount meaning? Yes. See from

$E(\theta) = \dfrac{\alpha}{\alpha+\beta}$ if $\alpha\uparrow, \beta\downarrow \Rightarrow \theta\uparrow \Rightarrow x\downarrow$

$\alpha\downarrow, \beta\uparrow \Rightarrow \theta\downarrow \Rightarrow x\uparrow$

α : pseudotrials witnessed

β : pseudofailures witnessed

$\theta \sim U(0,1)$ Laplace $\Rightarrow$ 1 pseudotrial witnessed, 1 pseudofailure $\Rightarrow E(\theta) = 0.5$

$\theta \sim \text{Beta}(0,0)$ Haldane $\Rightarrow$ nothing witnessed $E(\theta)$ d.n.e.

Why is $\hat{\theta}_{MLE}$?

$$\mathcal{L}(\theta;x) = p(x;\theta) = (1-\theta)^{\sum x_i}\theta^n$$

$$\ell(\theta;x) = \sum x_i \ln(1-\theta) + n\ln\theta$$

$$\ell'(\theta;x) = -\frac{\sum x_i}{1-\theta} + \frac{n}{\theta} \overset{set}{=} 0 \Rightarrow \frac{n}{\theta} = \frac{\sum x_i}{1-\theta} \Rightarrow \frac{1-\theta}{\theta} = \bar{x} \Rightarrow \frac{1}{\theta} - 1 = \bar{x} \Rightarrow \frac{1}{\theta} = \bar{x}+1 \Rightarrow$$

$$\hat{\theta}_{MLE} = \frac{1}{\bar{x}+1}$$

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

$$I(\theta) = E[-\ell''(\theta;x)] = E\left[\frac{\sum x_i}{(1-\theta)^2} + \frac{n}{\theta^2}\right] = \frac{\sum E(X_i)}{(1-\theta)^2} + \frac{n}{\theta^2} = \frac{nE(X_i)}{(1-\theta)^2} + \frac{n}{\theta^2} = n\left(\frac{E(X)}{(1-\theta)^2} + \frac{1}{\theta^2}\right)$$

$E(X) = \frac{1}{\theta} - 1$

$$= n\left(\frac{\frac{\cancel{1-\theta}}{\theta}}{(1-\theta)^{\cancel{2}}} + \frac{1}{\theta^2}\right) = n\left(\frac{1}{(1-\theta)\theta} + \frac{1}{\theta^2}\right)$$

$$= n\left(\frac{\theta}{(1-\theta)\theta^2} + \frac{1-\theta}{(1-\theta)\theta^2}\right) = n\left(\frac{1}{(1-\theta)\theta^2}\right)$$