Lec 9
Math 390 3/2/16

Recall

$P(\theta|x) \propto P(x|\theta)P(\theta)$ when $\theta \sim U(0,1)$

$P(\theta = 0.9 | x) \propto P(x | \theta = 0.9)(1)$

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

$$P_\theta(\theta)\left|\frac{d\theta}{do}\right|$$
$$\parallel$$
when $P(o) = P_\theta(t^{-1}(o)) \left|\frac{d}{do}[t^{-1}(o)]\right|$
$= (1)\frac{1}{(o+1)^2}$

$P(o|x) \propto P(x|o)\, P(o)$
$= P(x|o)\frac{1}{(o+1)^2}$

$o = \frac{0.9}{1-0.9} = 9$

$P(o = 9|x) \propto P(x|o=9)\frac{1}{(9+1)^2}$
$\parallel \qquad \parallel$
$P(\theta = 0.9|x) \propto P(x|\theta = 0.9) \cdot \underline{.01}$

off by $\frac{1}{100}$!

Why on earth should using odds matter??

This is the problem

What if there was a prior, j, that yielded same results regardless of scale?

$$P(\theta|x) \propto P(x|\theta)\, j(\theta)$$

$\forall\, t$ 1:1 of θ?

$$\|$$

$$P(\phi|x) \propto P(x|\phi)\, j(\phi) \qquad \forall x!$$

$$\|$$

$$P(t(\theta)|x) \propto P(x|t(\theta))\, j(t(\theta))$$

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

Is there a function that satisfies this??

$$f(y) = f(x)\left|\frac{dx}{dy}\right| \quad \text{where } y = t(x)$$

Find f! HARD!!!

Claim $j(\theta) \propto \sqrt{I(\theta)}$ is this function.

OK then, WTS $j(\phi) \propto \sqrt{I(\phi)}$

We know

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

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

where $\phi = t(\theta)$

$$I(\phi) = E[-\ell''(x;\phi)] = E\left[-\frac{d^2}{d\phi^2}\Big[\underbrace{\ln p(x|\phi)}_{g(\phi)}\Big]\right] \qquad \frac{d^2}{d\phi^2} g(\phi) =$$

let $\phi = t(\theta)$

At θ

$\ln p(x|\phi)$

$= \ln p(x|\theta)$

$$\frac{d}{d\phi^2}\left[g(t(\theta))\right]$$

$$\frac{d}{dx^2}\left[f(y)\right] = \frac{d}{dx}\left[f(y)\right]\frac{dy}{dx} = f''(y)\left(\frac{dy}{dx}\right)^2$$

$$= \frac{d^2}{dy^2}\left[f(y)\right]\left(\frac{dy}{dx}\right)^2$$

$$= \frac{d^2}{d\theta^2}\left[g(t(\theta))\right]\left(\frac{d\theta}{d\phi}\right)^2$$

$$= E\left[-\frac{d^2}{d\theta^2}\left[\ln p(x|\theta)\right]\left(\frac{d\theta}{d\phi}\right)^2\right]$$

$$= \underbrace{E\left[-\frac{d^2}{d\theta^2}\left[\ln p(x|\theta)\right]\right]}_{I(\theta)}\left(\frac{d\theta}{d\phi}\right)^2$$

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

Jeffreys prior for binomial lik.

Let's see this for the ex $\quad j(\theta) = \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$

$$p(\theta|x) \propto p(x|\theta)\, \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$$

$\theta = 0.9$

$\theta = \frac{o}{o+1}$

$\Rightarrow o = 9$

$$p(o|x) \propto p(x|o)\, j(o)$$

$$j(o) \propto \sqrt{I(o)} = \frac{}{\left(\frac{o}{o+1}\right)^{\frac{1}{2}}\left(1-\frac{o}{o+1}\right)^{\frac{1}{2}}}$$

Clear $\Rightarrow$ Same answer

Why not any old function eg $P(\theta) = \theta^2$

$P(\phi) = \left(\frac{\phi}{\phi+1}\right)^2$ Same answer!

But no! $P(\phi) \neq P(\theta)\frac{d\theta}{d\phi}$

So $P(\phi)$ and $P(\theta)$ do not represent same prob./ degree of prior belief... they are different models

$\phi = x$ $\quad y = g(x)$
$\theta = y$

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

$$\frac{d}{dx} f(y) = \frac{d}{dx} f(y) = \frac{d}{dy} f(y)\frac{dy}{dx} = f'(y)\frac{dy}{dx}$$

$$= \sqrt{E\left[-\frac{d^2}{d\theta^2}\left[\ln p(X|\theta)\right]\frac{d^2\theta}{d\phi^2}\right]}$$

$$\frac{d^2}{dx^2}\left[f(y)\right] = \frac{d}{dx}\left[f'(y)\frac{dy}{dx}\right] = f''(y)\frac{d^2y}{dx^2}$$

$$= \frac{d^2}{dy^2}\left[f(y)\right]\frac{d^2y}{dx^2}$$

$$= \sqrt{E\left[-\frac{d^2}{d\theta^2}\left[\ln p(X|\theta)\right]\right]}\left|\frac{d\theta}{d\phi}\right|$$

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

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

With lots of data, the x and the $n-x \to \infty$ as long as $\theta \in (0,1)$ ie $\neq 0$ or 1

It can be shown
$$\lim_{\alpha\to\infty} \frac{\text{Beta}(\alpha,\alpha) - \overset{\frac{1}{2}}{\frac{\alpha}{2\alpha}}}{\sqrt{\frac{\alpha^2}{(2\alpha)^2(2\alpha+1)}}} \to N(0,1)$$

$$\sqrt{\frac{1}{8\alpha+4}}$$

In fact
$$\frac{\theta|X - \overbrace{E[\theta|X]}^{\hat{\theta}_{MMSE}}}{SE[\theta|X]} \longrightarrow N(0,1)$$

$$\Rightarrow \theta|X \overset{\text{approx}}{\sim} N\left(E[\theta|X],\ SE[\theta|X]^2\right)$$

the Bayesian CLT

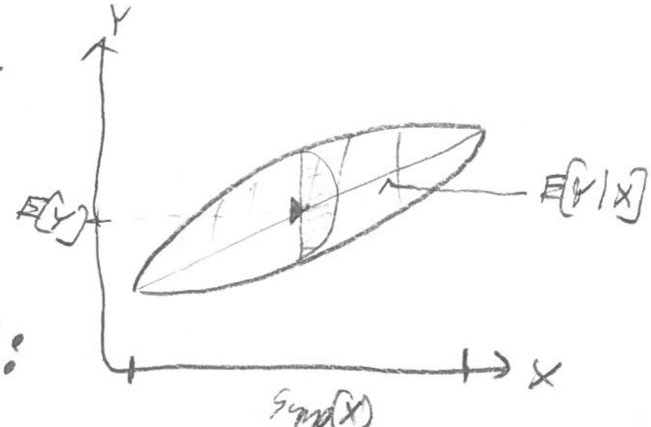
Only works if n is large. Why bother? We have computers and can calculate exactly. But the Z approx is useful sometimes

$\theta \longrightarrow \theta|X$ Bayesian Conditionalization

$CR_{\theta,1-\alpha} \supset CR_{\theta|X,1-\alpha}$? Makes sense that you get more sure, right?

Shouldn't $Var[\theta|X] < Var(\theta)$?

Need fact from prob theory...



Law of Iterated Expectation / Total Exp.

$$E_Y(Y) = E_X\Big(\underbrace{E_Y(Y|X)}\Big)$$

proof:

What I'm integrating w.r.t.

$$= E_X[\,g(X)\,]$$

$$= \int_{Supp(X)} E_Y(Y|X)\, p(x)\, dx \qquad \text{s.t. } E_Y(Y|X) = \int_{Supp(Y)} y\, p(y|x)\, dy$$

$$= \int_{Supp(X)} \left(\int_{Supp(Y)} y\, p(y|x)\, dy \right) p(x)\, dx = \int_{Supp(Y)} \left(\int_{Supp(X)} y\, p(y|x)\, p(x)\, dx \right) dy$$

$$= \int_{Supp(Y)} y \left(\int_{Supp(X)} p(x,y)\, dx \right) dy$$

$$= \int_{Supp(Y)} y\, p(y)\, dy = E(Y)$$

$$\operatorname{Var}_Y(Y) = E(Y^2) - (E(Y))^2$$

$$= E_X\left[E_Y[Y^2|X]\right] - \left(E_X\left[E_Y(Y|X]\right]\right)^2$$

Note $\operatorname{Var}_Y[Y|X] = E_Y[Y^2|X] - \left(E_Y(Y|X)\right)^2$

$$\Rightarrow E_Y(Y^2|X) = \operatorname{Var}_Y(Y|X] + \left(E_Y[Y|X]\right)^2$$

$$\to E_X\left[\operatorname{Var}_Y(Y|X] + \left(E_Y(Y|X)\right)^2\right] - \left(E_X\left[E_Y(Y|X]\right]\right)^2$$

Exp is lin op

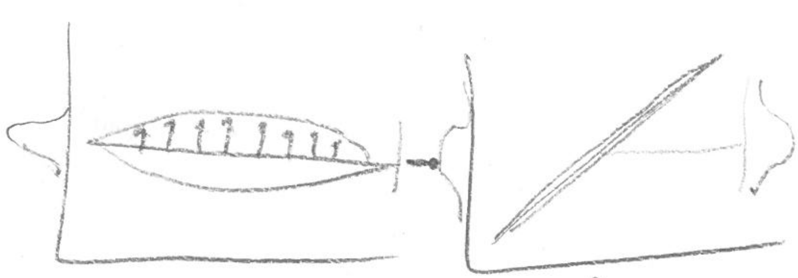
$$= E_X\left[\operatorname{Var}_Y(Y|X)\right] + E_X\left[\left(\underbrace{E_Y(Y|X)}_{Q(X)}\right)^2\right] - \left(E_X\left[\underbrace{E_Y(Y|X]}_{Q(X)}\right]\right)^2$$

group terms

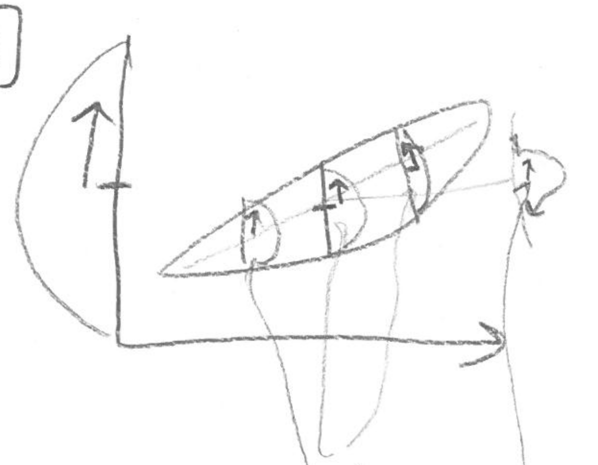
$$\to E_X(Q^2) - (E_X(Q))^2$$

$Y|X \sim N(\beta_0+\beta_1 x, \sigma^2)$

$$= E_X\left[\operatorname{Var}_Y(Y|X]\right) + \operatorname{Var}_X\left[E_Y(Y|X]\right]$$



e.g. $\operatorname{Var}_Y E_Y(Y|X) = 0$ e.g. $E_X \operatorname{Var}_Y(Y|X) = 0$



$E_X \operatorname{Var}_Y(Y|X) + \operatorname{Var}_X E_Y(Y|X)$

Use Formula, but let:

$Y = \theta, \ X = X$

$$\Rightarrow \operatorname{Var}_{\theta}(\theta) = \underset{X}{E}\left(\operatorname{Var}_{\theta}(\theta|X)\right) + \operatorname{Var}_{X}\left[\underset{\theta}{E}(\theta|X)\right)$$

$$\Rightarrow E\left[\operatorname{Var}(\theta|X)\right] = \underbrace{\operatorname{Var}[\theta] - E\left[\underbrace{\operatorname{Var}(\theta|X)}_{\geq 0}\right]}_{\geq 0}$$

$$\Rightarrow \underset{X}{E}\left[\operatorname{Var}_{\theta}(\theta|X)\right] < \operatorname{Var}_{\theta}(\theta)$$

We expect post. var to be less than prior variance.

e.g. $\theta \sim U(0,1)$

$\Rightarrow \operatorname{Var}[\theta] = \frac{1}{12}$

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

$Y \sim \text{Beta}(\alpha, \beta) \Rightarrow$
$\operatorname{Var}(Y) = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

$$\operatorname{Var}_{\theta}[\theta|x] = \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

let's say $n=1$

$$= \frac{(1+x)(2-x)}{\underset{\overset{\shortparallel}{3^2}}{9} \cdot 4} = \frac{2+x-x^2}{36}$$

$$E_X\left(\operatorname{Var}_{\theta}(\theta|x)\right) = E_X\left[\frac{2+x-x^2}{36}\right] = \frac{1}{18}\underset{\overset{\shortparallel}{\int p(x=1|\theta)p(\theta)d\theta}}{P(X=1)} + \frac{1}{18}\underset{\overset{\shortparallel}{\int p(x=0|\theta)p(\theta)d\theta}}{P(X=0)} = \frac{1}{18} < \frac{1}{12}$$

But this is only an expectation!

$$E_X\left[\mathrm{Var}_\theta(\theta|X)\right] < \mathrm{Var}_\theta(\theta)$$

For some X's it's possible that

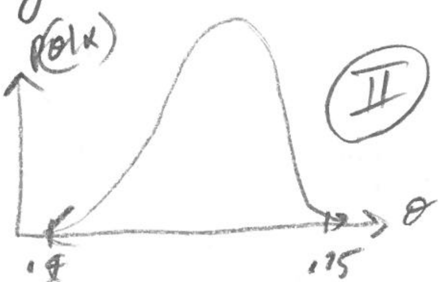
$$\mathrm{Var}_\theta[\theta|X] > \mathrm{Var}_\theta(\theta)$$

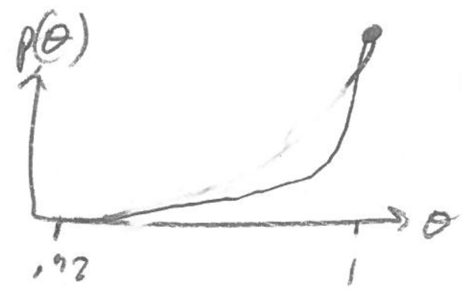
which is $\underline{\underline{\text{super}}}$ BAD!

$\Rightarrow$ It means you <u>failed</u> to reduce your uncertainty about θ!

When does this happen? When prior is <u>misspecified</u>.

e.g. $\theta \sim \mathrm{Beta}(100, 1) \Rightarrow$ I believe before the coin is unfair weighted towards heads





Now flip 78 times

get 6H, 72T

$$\mathrm{Var}[\theta] = \frac{(100)(1)}{(100+1)^2(100+1+1)} = .000096$$

$$\theta|X \sim \mathrm{Beta}(\alpha+x, \beta+n-x)$$
$$= \mathrm{Beta}(100+6, 1+72)$$
$$= \mathrm{Beta}(106, 73)$$

$$\mathrm{Var}(\theta|X) = \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

$\wedge$

$$\mathrm{Var}[\theta|x=6;n=78] = \frac{(106)(73)}{(179^2)(180)} = .001342$$

BAD!! why?

Prior was BAD!!!

What did we just do?

We looked at something which is based on prior, $P(\theta)$

then based on data, and asked: "is this reasonable?"

Check #1

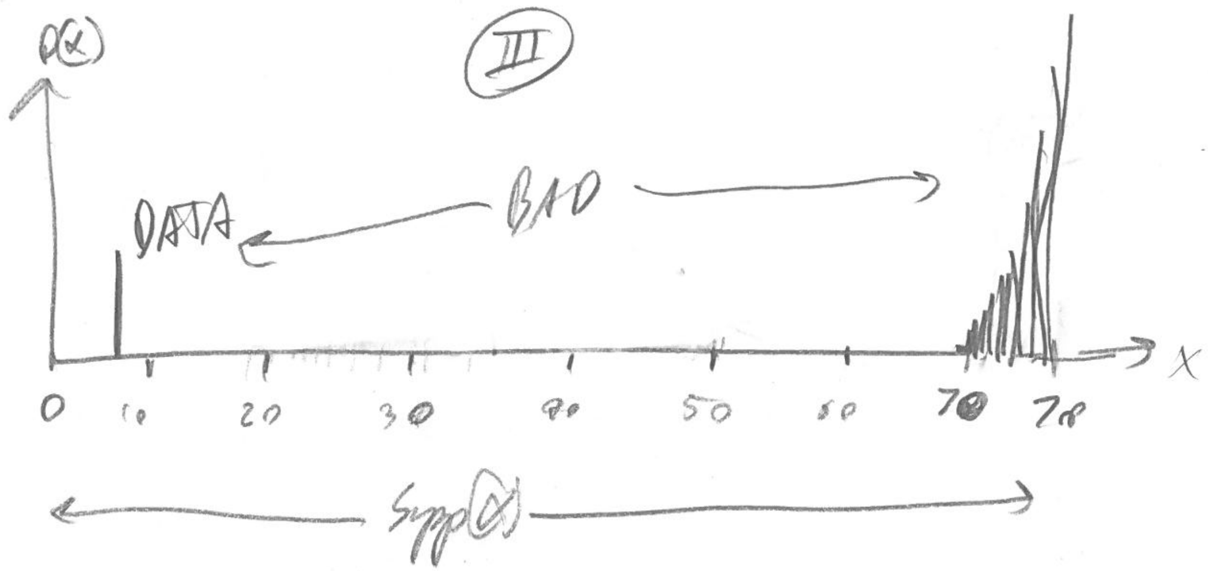
If $Var[\theta] < Var[\theta|X]$ ... something may be wrong

Check #2

$$P(X) = \int_{\Theta} P(X|\theta)P(\theta)d\theta$$

marginal likelihood, denominator in Bayes Rule, const of proportionality, "prior predictive dist" (HW 2)

$X \sim BetaBin(n, \alpha, \beta) = BetaBin(78, 100, 1)$ which looks like



Look above! X is very not probable considering P(X)!

prior predictive check

If $\theta \sim U(0,1) \Rightarrow X \sim \text{BetaBin}(78, 1, 1)$

DATA is reasonable given $P(X)$

$P(X)$ should be a fairly diffuse distr. unless you have serious prior information.

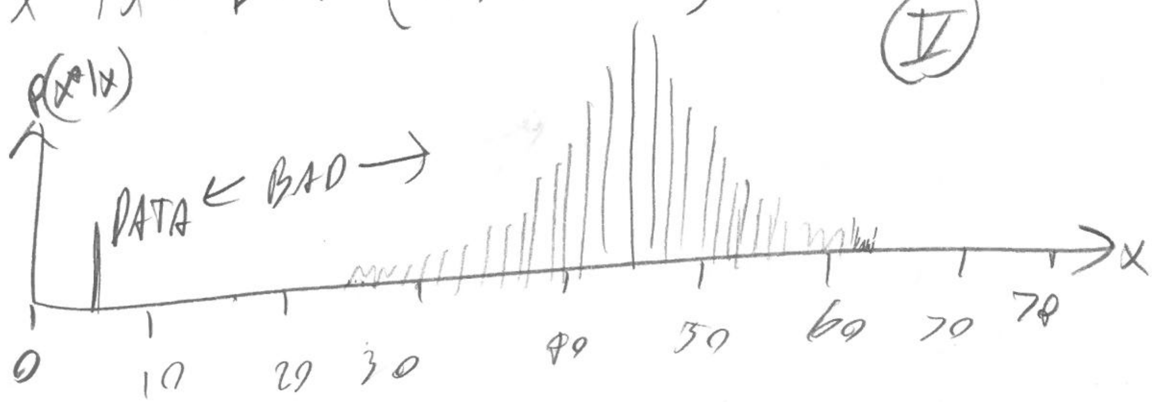
"X^{rep}"

$X^{rep} := X^*$ with $m=n$

Check #3

The data should look plausible under $P(X^* | X)$, the posterior predictive distr. using $m=n$. Same sample size!

Here $X^* | X \sim \text{BetaBin}(\overset{n = m}{78}, \overset{x+\alpha}{106}, \overset{\beta+n-x}{73})$



If $\theta \sim U(0,1)$

$\Rightarrow X^* | X \sim \text{BetaBin}(78, 7, 73)$

VI

GOOD!

Usually do check #3. If fails, do #1 & #2.

Checks 1, 2, 3 are "self-consistency" checks. If the likelihood and prior truly model $P(X,\theta)$, then X should be reasonable before and after! Check #3 is an "external validation." Checks 2 & 3 are "graphical checks."

Is our model true or false? Always FALSE in an absolute sense. George Box: "All models are wrong but some are useful."

Bad question! Better question: "do the model's deficiency have a noticeable impact on inference?"

If model is deficient $\Rightarrow$ inference and prediction will be BAD!