Lecture 9 Math 341 3/9/17

Recall $X|\theta \sim \text{Bin}(n, \theta)$, n fixed, known

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

$\theta|X \sim \text{Beta}(\underbrace{\alpha+x}_{\alpha'}, \underbrace{\beta+n-x}_{\beta'})$ ← conjugacy

$X^*|X \sim \text{BetaBinom}(n^*, \underbrace{\alpha+x+x^*}_{\alpha'}, \underbrace{\beta+n-x+n^*-x^*}_{\beta'}) := \binom{n^*}{x^*} \frac{B(\alpha+x+x^*, \beta+n-x+n^*-x^*)}{B(\alpha+x, \beta+n-x)}$

← posterior α, β

if $n^* = 1$

$X^*|X \sim \text{BetaBinom}(1, \alpha+x+x^*, \beta+n-x+n^*-x^*) = \frac{B(\alpha+x+x^*, \beta+n-x+n^*-x^*)}{B(\alpha+x, \beta+n-x)} = \text{Bern}\left(\frac{\alpha+x}{\alpha+\beta+n}\right)$

we know this is Bernoulli... so set $x^* = 1$, find prob.

$$P(x^*=1 \mid x) = \frac{B(\alpha+x+1, \beta+n-x)}{B(\alpha+x, \beta+n-x)} = \frac{\frac{\Gamma(\alpha+x+1)\,\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n+1)}}{\frac{\Gamma(\alpha+x)\,\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}} = \frac{(\alpha+x)}{(\alpha+\beta+n)}$$

$$= \int_\Theta P(x|\theta)\, P(\theta|x)\, d\theta$$

$P(x^*|x)$ is dist of future x^* given data "posterior predictive dist"

$P(x)$ = dist of data observed here it could be thought of as

$$= \int_\Theta P(x|\theta)\, P(\theta)\, d\theta$$

$P(x|\{\})$

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

$= P(\theta)$

⇒ same form!

$X \sim \text{BetaBinom}(n, \alpha+x, \beta+n-x)$

AKA the "prior predictive dist"

$P(\theta)$ ... this denominator has another name!

If conjugate, $P(\theta)$, $P(\theta^*|x)$ belong to same family as well. Why?

Uniform prior: a prior which does not have a "large" effect on posterior

$\theta \sim U(0,1) = \text{Beta}(1,1)$

$\theta|x \sim \text{Beta}(\overset{\alpha}{1}+x, \overset{\beta}{1}+n-x) \Rightarrow \hat{\theta}_{mse} = \frac{x+1}{n+2}$ AKA Laplace estimate

$U(0,1)$ can thought "influence" yielded pseudocounts of 1 success, 1 failure. That's not "no information".

What would NO info look like? Uniform? YES

$\theta|x \sim \text{Beta}(\overset{\alpha}{0}+x, \overset{\beta}{0}+n-x) \Rightarrow \hat{\theta}_{mse} = \frac{x}{n} = \hat{\theta}_{MLE}$

$\Rightarrow \theta \sim \text{Beta}(0,0)$ what's wrong? $\alpha > 0, \beta > 0$

this is an illegal prior! AKA "improper prior"

But posterior is proper if $x \neq 0$ and $x \neq n$

Tons of theory on this... some say improper priors okay, some say no. Okay for me... except... you must be careful your posterior is proper!

$\theta \sim \text{Beta}(1,1) \Rightarrow$ indifference but successes and failures are known to be possible

$\theta \sim \text{Beta}(0,0) \Rightarrow$ successes and failures not known to be possible

AKA complete ignorance

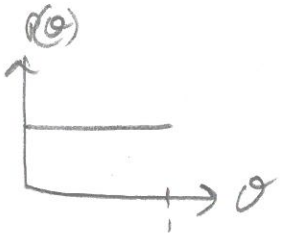
Haldane prior (1932)

Jupiter problem

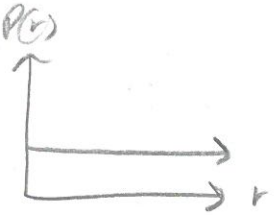
$\theta \sim U(0,1)$ means every prob equally likely

What if I consider the odds, $r := \frac{\theta}{1-\theta}$. Am I indifferent on this scale?

No... not possible!



$\theta \in (0,1)$



$r \in (0,\infty)$

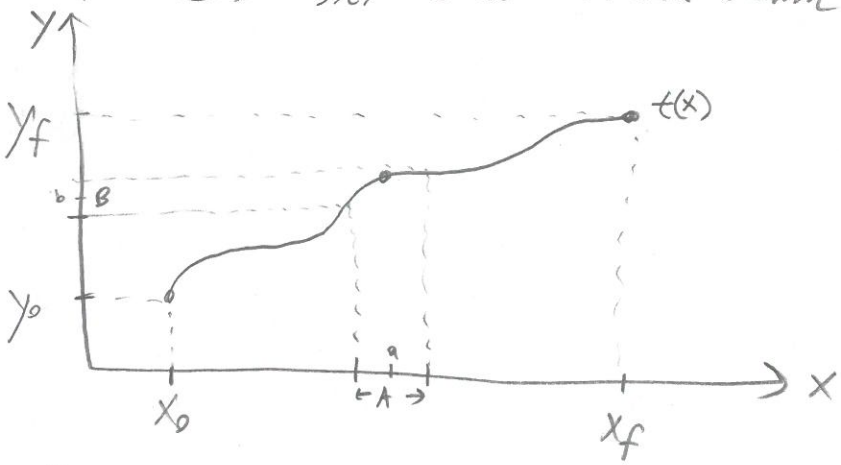
the principle of indifference has a problem!

What is PDF of R?

Math 621 covers transformation of variables.

Imagine r.v.'s X, Y with densities f_X, f_Y such f_X known, f_Y unknown

Suppose $Y = t(X)$ s.t. t is an invertible function



$Supp[X] = [x_0, x_f]$, $Supp[Y] = [y_0, y_f]$

$P(X \in A) \approx f_X(a) A$

$P(Y \in B) \approx f_Y(b) B$

$P(X \in A) = P(Y \in B)$

$\Rightarrow f_X(a) A = f_Y(b) B$ Note $A > 0, b > 0$

let a be small

$f_X(a)|dx| = f_Y(b)|dy|$ s.t. $b = t(a)$ or $a = t^{-1}(b)$

Solve for $f_Y(y)$ $\Rightarrow f_Y(b) = f_X(a)\left|\frac{dx}{dy}\right|$

let $b = y$ the usual dummy variable

$\Rightarrow f_Y(y) = f_X\left(t^{-1}(y)\right)\left|\frac{d}{dy}\left[t^{-1}(y)\right]\right|$

What is PDF of R? $\theta \sim U(0,1) \Rightarrow f_\theta(\theta) = 1$ $\quad R - RX = X \Rightarrow R = X + RX \Rightarrow R = (1-R)X \Rightarrow$

$$R = t(\theta) = \frac{\theta}{1-\theta} \quad \Rightarrow \theta = t^{-1}(R) = \frac{R}{R+1}$$ quotient rule

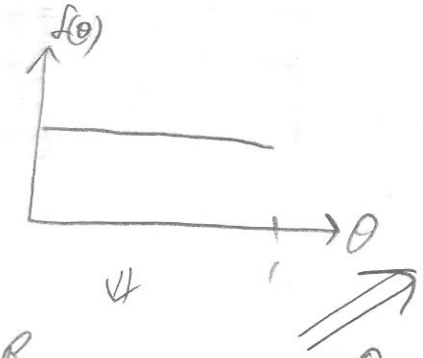
reparameterization!

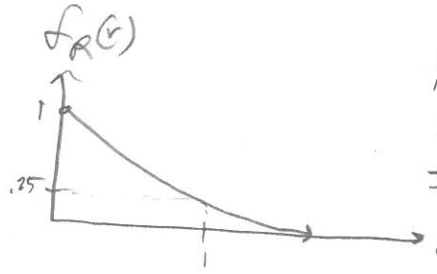
$$f_R(r) = f_X\left(t^{-1}(r)\right)\left|\frac{d}{dr}\left[t^{-1}(r)\right]\right| = \left|\frac{r(1)-(r+1)}{(r+1)^2}\right| = \frac{1}{(r+1)^2}$$

X = θ

is this a density? $\int_{Supp(R)} f_R(r)\,dr = 1 \Rightarrow \int_0^\infty \frac{1}{(r+1)^2}\,dr = \left[\frac{r}{r+1}\right]_0^\infty = 1$ ✓

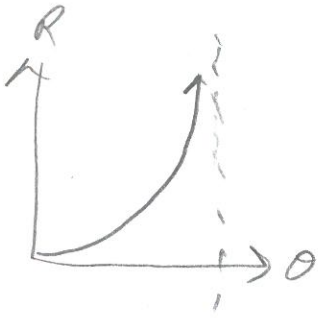
Now $f_R(r) \neq f_\theta(\theta)$ ⇒ If you are indifferent about θ, you are <u>not</u> indifferent about odds or any other monotone transformation of θ





Fisher used this to show Bayes is stupid!

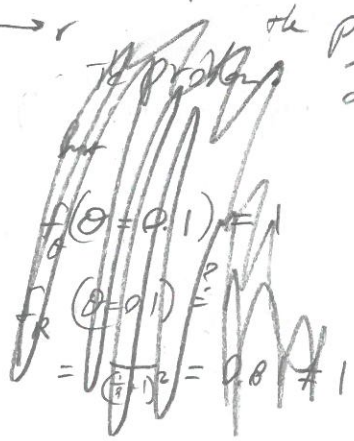
Is this a problem if the prior is discrete?



$\theta \in [0, 0.5] \Rightarrow R \in [0,1]$

$\theta \in [0.5, 1] \Rightarrow R \in [1, \infty)$

$r = \frac{0.1}{1-0.1} = \frac{1}{9}$



Big question: is there any way out of this?

Is there a prior θ or $\theta(\theta)$ s.t. any reparameterization will yield the prior density?

Beta(1,1)
"
$\theta \sim U(0,1)$ Laplace prior
$\theta \sim \text{Beta}(0,0)$ Haldane prior

} Uniform, Vague, Weak, Diffuse

⇓

don't affect inference too much.

Strategy
protocol

Is there a way to choose any uniform prior that would be the same under reparameterizations?

Likelihood Model

$p(x|\theta) \xrightarrow{\text{strategy}}$ pick $p(\theta)$ and make a reparameterization:

$\phi = t(\theta)$: s.t. t is 1:1 and monotone

$p(x|\phi) \xrightarrow{\text{strategy}}$ pick $p(\phi)$

Wouldn't it be nice if

$$p(\phi) = p\left(t^{-1}(\phi)\right)\left|\frac{d}{d\phi}\left[t^{-1}(\phi)\right]\right|$$

As is, the strategy itself would ensure this "invariance" wouldn't break?

This is the strategy Jeffreys invented $\approx$ 1930's.

Before we get there, we need two pieces.

(1) "Kernels"

(2) Fisher Information

Kernels Recall ...

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{p(x)} \propto p(x|\theta)\,p(\theta)$$ why? $p(x)$ was a function of θ!

$$f(x;\theta) \propto g(x;\theta)$$

this means by def that $\exists\, c \in \mathbb{R}$ s.t.

$$f(x;\theta) = \frac{1}{c}\, g(x;\theta)$$ where $\frac{1}{c}$ is called $h(\theta)$

How to find c?

Note: $\frac{f(x_1;\theta)}{f(x_2;\theta)} = \frac{h(\theta)}{h(\theta)} \frac{g(x_1;\theta)}{g(x_2;\theta)}$

Note: $\int_{\mathrm{supp}(x)} f(x)\,dx = 1$

$$\int_{\mathrm{supp}(x)} g(x)\,dx = \int c f(x)\,dx = c \underbrace{\int f(x)\,dx}_{1} = c = \int g(x)\,dx$$

Note: $\int g(x)\,dx < \infty$ ... and ... $\int g(x)\,dx > 0$ Note: $g(x)$, $f(x)$ are 1:1

r.v.'s can be identified by their kernels

So let's see it:

kernel

$$p(\theta|x) \propto \left(\binom{n}{x}\theta^x(1-\theta)^{n-x}\right)\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

$$\propto \mathrm{beta}(x+\alpha, n-x+\beta)$$

why? $\theta \sim \mathrm{Beta}(\alpha,\beta) := \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} = \underbrace{\theta^{9}(1-\theta)^{6}}$

the kernel of the beta

$$X|\theta \sim \mathrm{Bin}(n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^n(1-\theta)^{-x}$$

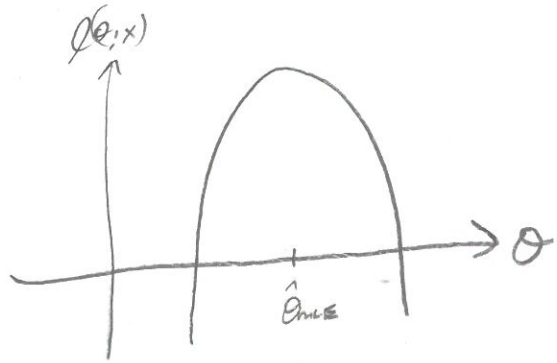
$$\propto \frac{1}{x!(n-x)!}\left(\frac{\theta}{1-\theta}\right)^x$$ ← kernel of the binomial!!

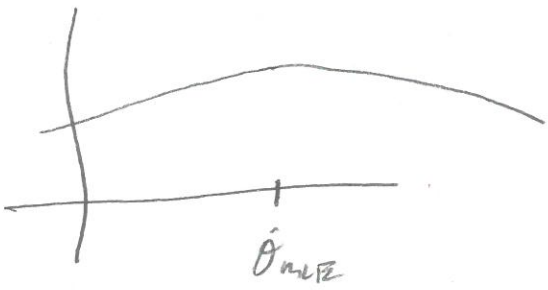
If you know a scientific value $p(\theta|x) \propto$ kernel ... you're done! EASIER than solving explicitly.

Fisher Info

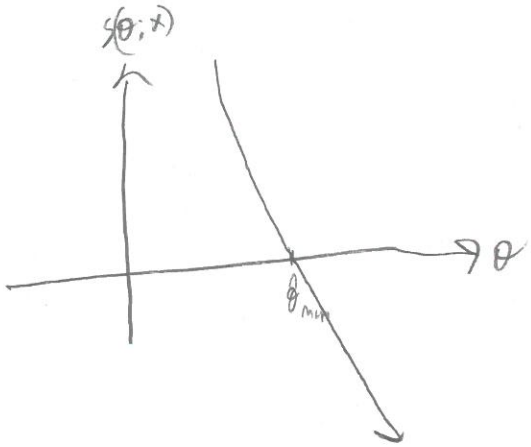
Recall likelihood and log likelihood

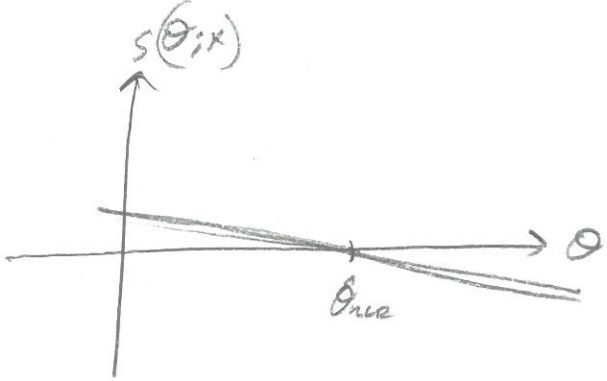
$$\mathcal{L}(\theta; x) = p(x; \theta) \qquad \ell(\theta; x) := \ln\left(\mathcal{L}(\theta; x)\right)$$



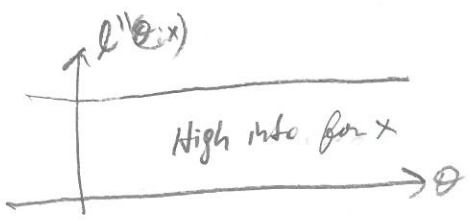


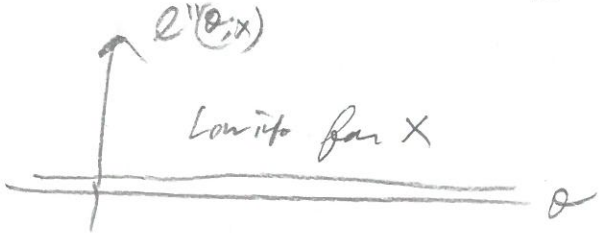
Define $s(\theta; x) := \ell'(\theta; x)$





$$I(\theta) := \operatorname{Var}_x\left[s(\theta; x)\right] \overset{\text{Stat 633}}{=} \ldots = \mathbb{E}_x\left[s(\theta; x)^2\right] \overset{\text{Stat 633}}{=} \ldots = \mathbb{E}_x\left[-\ell''(\theta; x)\right]$$





$I(\theta)$ 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

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

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

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

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

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

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

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

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

If $\theta = \frac{1}{2}$, $n = 1$, How much info? $I\left(\frac{1}{2}\right) = 4$ ← 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 nice, function of n?

More data ⇒ more info. Recall binomial is n bernoullis... more bernoulli data ⇒ more info

---

Back to the issue... CONSIDER:

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