2/21/17 Lec 6 Math 341

$\mathcal{F}$ = Binomial w/ fixed n recall...

$$\theta \sim \text{Beta}(\alpha, \beta) := \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}, \quad E[\theta] = \frac{\alpha}{\alpha+\beta}, \quad Var[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$P(\theta \mid x) = \frac{P(X \mid \theta) P(\theta)}{\underbrace{P(X)}_{\int_{\Theta} P(X|\theta) P(\theta)\, d\theta}} = \frac{\cancel{\binom{n}{x}} \theta^x (1-\theta)^{n-x} \cancel{\frac{1}{B(\alpha,\beta)}} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{\int_0^1 \cancel{\binom{n}{x}} \theta^x (1-\theta)^{n-x} \cancel{\frac{1}{B(\alpha,\beta)}} \theta^{\alpha-1} (1-\theta)^{\beta-1}\, d\theta} = \frac{\theta^{x+\alpha-1} (1-\theta)^{n-x+\beta-1}}{\int_0^1 \theta^{x+\alpha-1} (1-\theta)^{n-x+\beta-1}\, d\theta}$$

$$= \frac{1}{B(x+\alpha, n-x+\beta)} \theta^{x+\alpha-1} (1-\theta)^{n-x+\beta-1} = \text{Beta}(x+\alpha, n-x+\beta)$$

$$\underset{\text{Beta}(\alpha,\beta)}{\theta} \xrightarrow{\ X\ } \underset{\text{Beta}(x+\alpha,\, n-x+\beta)}{\theta \mid x}$$

The Beta is the conjugate prior for the Binomial.

Posterior pt estimation

$$\hat{\theta}_{MMSE} = E[\theta \mid x] = \frac{x+\alpha}{n+\alpha+\beta}$$

$$\hat{\theta}_{MAP} = \text{Mode}[\theta \mid x] = \frac{x+\alpha-1}{n+\alpha+\beta-2} \quad \text{if } x+\alpha > 1 \ \& \ n-x+\beta > 1$$

$$\hat{\theta}_{MAE} = \text{Med}[\theta \mid x] = \text{qbeta}(0.5, x+\alpha, n-x+\beta)$$

no closed form sol'n

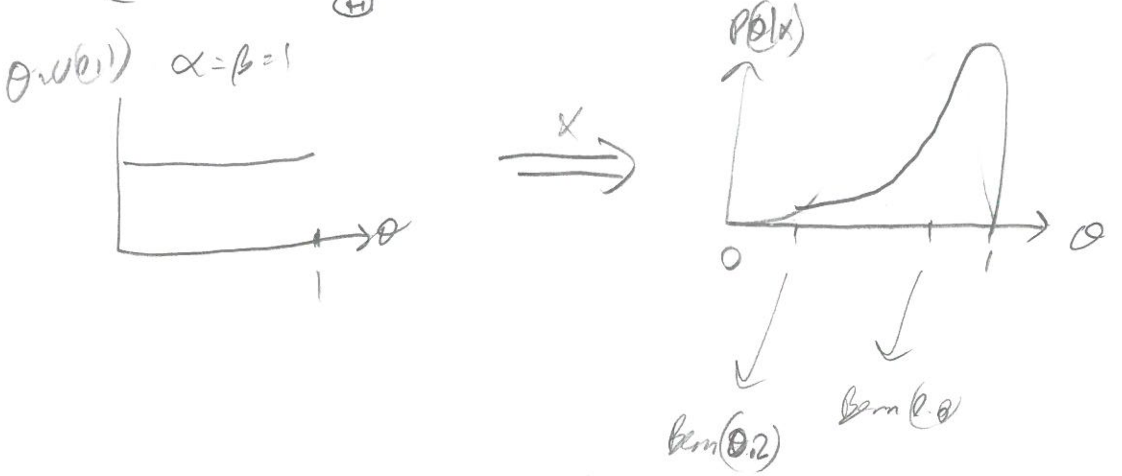
Now let's ask the question... the next X^* ... what is its distr?

$n^* = 1$ (one trial)

We know it must be Bernoulli since $\mathrm{Supp}[X^*] = \{0,1\}$

$$P(X^*|X) = \int_{\Theta} P(X^*|\theta)\, P(\theta|X)\, d\theta \qquad \text{same as} \qquad f(x|z) = \int_{\mathrm{Supp}(Y)} f(x,y|z)\, dy$$

$$f(x,y|z) = f(x|y,z)\, f(y|z)$$



Draw out many, many θ's (prop. to how likely these θ's are after seeing the data) and avg them

$$P(X^*|X) = \int_0^1 \left(\theta^{X^*}(1-\theta)^{1-X^*}\right) \frac{1}{B(X+\alpha,\, n-X+\beta)}\, \theta^{X+\alpha-1}(1-\theta)^{n-X+\beta-1}\, d\theta$$

$$= \frac{1}{B(X+\alpha,\, n-X+\beta)} \int_0^1 \theta^{X^*+X+\alpha-1}(1-\theta)^{-X^*+n-X+\beta}\, d\theta$$

$$= \frac{B(X^*+X+\alpha,\, -X^*+n-X+\beta+1)}{B(X+\alpha,\, n-X+\beta)}$$

this is the answer... but we want to write it as a Bernoulli

Recall...
$\Gamma(x+1) = x\Gamma(x)$

$$P(X^*=1|X) = \frac{B(1+X+\alpha,\, n-X+\beta)}{B(X+\alpha,\, n-X+\beta)} = \frac{\dfrac{\Gamma(1+X+\alpha)\,\cancel{\Gamma(n-X+\beta)}}{\Gamma(1+n+\alpha+\beta)}}{\dfrac{\Gamma(X+\alpha)\,\cancel{\Gamma(n-X+\beta)}}{\Gamma(n+\alpha+\beta)}} = \frac{\dfrac{(X+\alpha)\,\cancel{\Gamma(X+\alpha)}}{(n+\alpha+\beta)\,\cancel{\Gamma(n+\alpha+\beta)}}}{\dfrac{\cancel{\Gamma(X+\alpha)}}{\cancel{\Gamma(n+\alpha+\beta)}}} = \frac{X+\alpha}{n+\alpha+\beta}$$

$$\Rightarrow X^*|X \sim \mathrm{Bern}\left(\frac{X+\alpha}{n+\alpha+\beta}\right)$$

$$\uparrow$$

$$E[\theta|X]$$

Why does this make sense? The avg of many draws will be the average $\theta|X$!

Let's take a look at the posterior again

$$\text{Beta}(\alpha,\beta) \xrightarrow{\;x\;} \text{Beta}(\alpha + x,\ \beta + n - x)$$

α: # of prior successes (AKA "pseudosuccesses")

x: # of data successes

β: # of prior failures (AKA "pseudofailures")

$n - x$: # of data failures

α, β are "pseudocounts"

Conjugate prior parameters have interpretations as pseudodata.
They are as if you've seen data before!

$\theta \sim U(0,1) = \text{Beta}(1,1) \Rightarrow$ as if you've seen $\alpha = 1$ success previously
and " " " " " $\beta = 1$ failure previously

$\theta \sim U(0,1)$ is <u>not</u> devoid of information. The principle of indifference
<u>is a statement</u> about a belief!

$$E(\theta) = \frac{\alpha}{\alpha+\beta} = \frac{1}{1+1} = 0.5 \Rightarrow$$ You believe that your prior uncertainty is centered around 0.5 (the prior expectation)

$\hat{\theta}_{MMSE} = E(\theta \mid x) = \dfrac{\alpha + x}{n + \alpha + \beta}$. Are $E(\theta)$ and $E(\theta \mid x)$ related??

ρ

$E[\theta]$ $\qquad \bar{X} = \hat{\theta}_{MLE}$

$$\hat{\theta}_{MMSE} = E[\theta|x] = \frac{x+\alpha}{n+\alpha+\beta} = \frac{\alpha+\beta}{\alpha+\beta}\cdot\frac{\alpha}{n+\alpha+\beta} + \frac{x}{n+\alpha+\beta}\cdot\frac{n}{n}$$

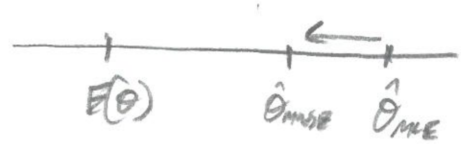
$$= \frac{\alpha+\beta}{n+\alpha+\beta}E[\theta] + \frac{n}{n+\alpha+\beta}\hat{\theta}_{MLE}$$

Note $\underbrace{\frac{\alpha+\beta}{n+\alpha+\beta}}_{\rho} + \underbrace{\frac{n}{n+\alpha+\beta}}_{1-\rho} = 1$

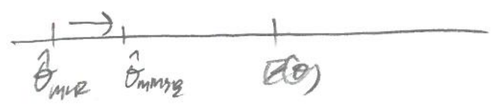
$$\hat{\theta}_{MMSE} = \rho E[\theta] + (1-\rho)\hat{\theta}_{MLE}$$

this is known as a "shrinkage estimator" because it shrinks the purely data-driven estimate towards the prior mean.

ρ is the shrinkage proportion



shrink down



shrink up

Large value of $\rho = \frac{\alpha+\beta}{n+\alpha+\beta}$ shrink "harder"

If α, β large compared to n ... $\Rightarrow$ "strong" prior or small sample size

" small " " " " " $n \Rightarrow$ "weak" prior or large sample size

$n\uparrow \Rightarrow \rho\downarrow$

$\lim_{n\to\infty} \rho = 0$. In the limit, large sample sizes "drown out" the prior.

Goals of Inference

① Pt est for θ ✓

  (1b) Prediction for future ✓ (for 1 obs)

② Conf. interval for θ ←

③ Testing claims of θ

Let's review Math 241

$n = 2$
$x = 1$ $\quad \hat{\theta}_{MLE} = 0.5$

$$CI_{\theta, 1-\alpha} := \left[\hat{\theta}_{MLE} \pm z_{\frac{\alpha}{2}} \, SE[\hat{\theta}_{MLE}] \right] \qquad SE[\hat{\theta}_{MLE}] = \sqrt{\frac{\theta(1-\theta)}{n}} \approx \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

many PhD theses have been written about this

$$CI_{\theta, 95\%} := \left[0.5 \pm 2 \sqrt{\frac{0.5 \cdot (1-0.5)}{2}} \right]$$
$$= [0.5 \pm 0.707] = [-0.21, 1.21]$$

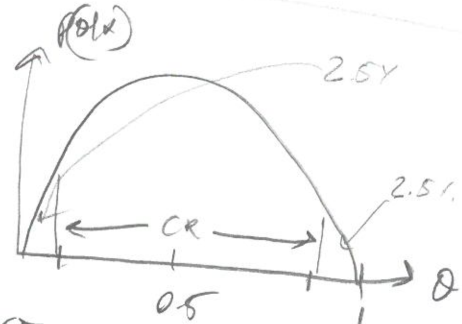
absurd! $\theta \in (0,1)$
At best completely useless!!!
Also.. can't get $P(\theta \in CI) = 95\%$.!!

Can we do better with Bayes?

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

as an example $\alpha = \beta = 1$ (uniform prior)

$\Rightarrow \theta | x \sim \text{Beta}(1+1, 1+(2-1)) = \text{Beta}(2,2)$



Now we want a credible region (CR) i.e. the Bayesian CI to have...

$P(\theta \in CR) = 1 - \alpha$

How to make CR??

$$CR_{\theta, 1-\alpha} := \left[\text{Quantile}[\theta|x, \tfrac{\alpha}{2}], \text{Quantile}[\theta|x, 1-\tfrac{\alpha}{2}] \right]$$

In our example,

$$CR_{\theta, 95\%} = \left[\text{Quantile}[\text{Beta}(2,2), 2.5\%], \text{Quantile}[\text{Beta}(2,2), 97.5\%] \right]$$
$$= [\text{qbeta}(.025, 2, 2), \text{qbeta}(.975, 2, 2)]$$
$$= [.094, .906]$$

much better!!!

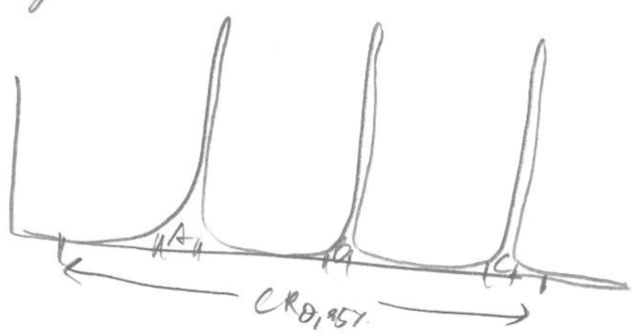
Why is a computer needed here?

$$\text{Quantile}\left[\text{Beta}(\alpha,\beta),\ 2.5\%\right] = x \text{ s.t. } \frac{1}{B(\alpha,\beta)}\int_0^x \theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta = .025$$

must be done numerically except in special cases of α, β.

Can we make a smaller CR? Yes...

Imagine



What about $A \cup B \cup C$. This has the smallest total length of the interval. It is called the highest density region (HDR) approach.

~~① Disadvantages having a non-contiguous CR... doesn't sit right with many people...~~
~~② Very computationally intensive to get it exact.~~
~~$\Rightarrow$ Other definition, the contiguous CR, we will be using~~

Disadvantages

① Not palatable to have a non-contiguous interval

e.g. $CR_{\theta,95\%} = [0.1, 0.2] \cup [0.8, 0.9]$ i.e. "low or high" makes little sense

② Very computationally intensive to find this region

$\Rightarrow$ we will use the contiguous CR defined above