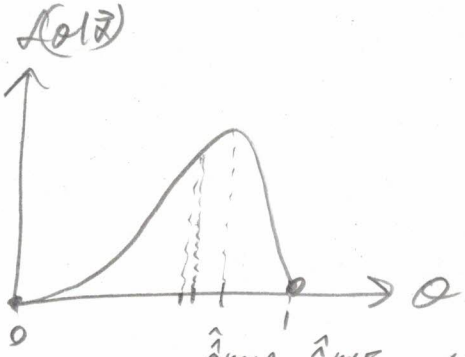
Lec 17 MATH 341/641

DGP: $X_1, \dots X_n \overset{iid}{\sim} Bern(\theta)$, $f(\theta) = U(0,1)$, Laplaces Prior of Indifference

$\vec{x} = \langle 0,1,1 \rangle \Rightarrow f(\theta|\vec{x}) = Beta(3,2)$



$\hat{\theta}^{MAP} = \hat{\theta}^{MLE} = .667$ (Mode)

$\hat{\theta}^{MSE} = 0.6$ (mean)

$\hat{\theta}^{MMAE} = .618$ (median)

Is there another reasonable point estimate?

Let $\hat{\theta}^{MMAE} := MED[\theta|\vec{x}]$ which is the $\hat{\theta}$ which minimizes $E[|\theta - \hat{\theta}| \,|\, \vec{x}]$

MMAE: "minimum mean absolute error" see next page

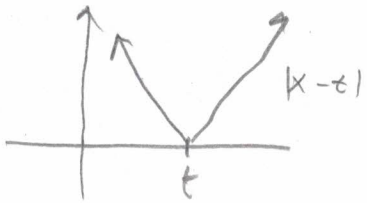
There is no closed form formula for median of the Beta dist so...

$$qbeta(q, \alpha, \beta) := \left\{ x : \int_0^x \frac{1}{B(\alpha,\beta)} \eta^{\alpha-1}(1-\eta)^{\beta-1} d\alpha = I_x(\alpha,\beta) = q \right\}$$

$$\hat{\theta}^{MMAE} = qbeta(0.5, \alpha, \beta) = qbeta(0.5, 3, 2) = .618$$

If $\vec{x} = \langle 0,0,1 \rangle$ $\hat{\theta}^{MMAE} = qbeta(0.5, 2, 1) = .159$, also reasonable for that scenario

This completes the discussion of point estimation in Bayesian Inference: $\hat{\theta}^{MAP}$, $\hat{\theta}^{MSE}$, $\hat{\theta}^{MMAE}$ these are 3 widely used estimators



Consider rv X , continuous

let $g(t) = E[|X-t|]$

$$g'(t) = \frac{d}{dt}[E[|X-t|]] = E\left[\frac{d}{dt}[|X-t|]\right] = E[\mathbb{1}_{X \geq t} - \mathbb{1}_{X<t}]$$

$$= E[\mathbb{1}_{X \geq t}] - E[\mathbb{1}_{X<t}] = P(X \geq t) - P(X > t) \overset{set}{=} 0$$

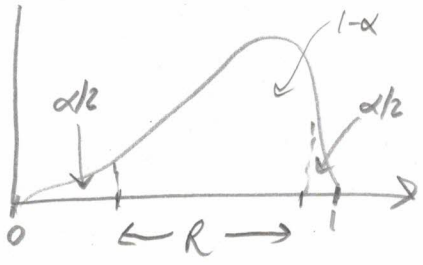
$$\Rightarrow P(X \geq t) = P(X > t)$$

$$\Rightarrow P(X \geq t) = 0.5 \ \& \ P(X > t) = 0.5$$

$$\Rightarrow t = MED[X]$$

Let's now do goal #2: confidence sets. Now, we get what we want. Let $R \subset \Theta$, a "region".

Get $P(\theta \in R \mid \vec{x}) = 1-\alpha$! How?



This is known as a 2-sided credible region:

$$CR_{\theta,1-\alpha} := \left[Q[\theta \mid \vec{x}, \tfrac{\alpha}{2}],\ Q[\theta \mid \vec{x}, 1-\tfrac{\alpha}{2}]\right]$$

A 95% CR would then be

$$CR_{\theta,95\%} = \left[Q[\theta \mid \vec{x}, 2.5\%],\ Q[\theta \mid \vec{x}, 97.5\%]\right]$$

In our example $\vec{x} = \langle 0,1,1\rangle$, $P(\theta) = U(0,1)$,

$$CR_{\theta,95\%} = [\text{qbeta}(.025,3,2),\ \text{qbeta}(.975,3,2)]$$

$$= [0.194,\ 0.932]$$ Makes sense! Why so wide? $n=3$

For $\vec{x} = \langle 0,0,0\rangle$

$$CR_{\theta,95\%} = [\text{qbeta}(.025,1,4),\ \text{qbeta}(.975,1,4)] = [0.006,\ 0.602]$$

Now let's do goal #3: hypothesis testing. Again we get what we want! We can compute $P(H_0|\vec{x})$, $P(H_a|\vec{x})$.

Define Bayesian p-value as $p_{vl} := P(H_0|\vec{x})$

Let's do one-sided tests first

$H_a: \theta > 0.5 \Rightarrow H_0: \theta \leq 0.5$ right-tailed test

$$p_{vl} = P(H_0|\vec{x}) = P(\theta \leq 0.5|\vec{x}) = \int_0^{0.5} \text{Beta}(3,2)\,d\theta = I_{0.5}(3,2)$$

regularized, incomplete beta function

In this class, use the R code notation $pbeta(x,\alpha,\beta) = I_x(\alpha,\beta)$

$p_{vl} = pbeta(0.5,3,2) = .3125 \Rightarrow$ Retain H_0. Possible Type II error since underpowered at $n=3$.
at $\alpha = 5\%$

Likewise the left-tailed test is...

$H_a: \theta < 0.5 \Rightarrow H_0: \theta \geq 0.5$

$p_{vl} = P(H_0|\vec{x}) = P(\theta \geq 0.5|\vec{x}) = 1 - P(\theta \leq 0.5|\vec{x}) = 1 - pbeta(0.5,3,2) = .6875$

$\Rightarrow$ Retain H_0.
at $\alpha = 5\%$

Okay let's do 2-sided tests:

$H_a: \theta \neq 0.5 \Rightarrow H_0: \theta = 0.5$

$p_{vl} = P(H_0|\vec{x}) = P(\theta = 0.5|\vec{x}) = 0$! Since $\theta|\vec{x}$ is a continuous dist.
In fact 0.5 is any $H_0: \theta = \theta_0$ will have $p_{vl} = 0$

What happened here? Exactly what should happen. Recall in frequentist hypothesis testing of $H_0: \theta = \theta_0$. If $n \to \infty \Rightarrow p_H \to 0$ and H_0 is always rejected? But the rejection may not be "practically significant" or "clinically significant". Who cares if $\theta = 0.500001$ for a coin? It's still fair for all "practical purposes"! The Bayesian Framework forces you to make clear what practical significance means by forcing you to define δ, a margin of equivalence. The hypotheses then become:

$$H_a: \theta \notin [\theta_0 \pm \delta] \Rightarrow H_0: \theta \in [\theta_0 \pm \delta]$$

Let $\delta = 0.02$

$$p_H = P(H_0 | \vec{x}) = P(\theta \in [0.5 \pm .02] | \vec{x}) = P(\theta \leq .52 | \vec{x}) - P(\theta \leq .48 | \vec{x})$$

$$= pbeta(.52, 3, 2) - pbeta(.48, 3, 2) = .060 \Rightarrow \text{Retain } H_0 \text{ at } \alpha = 5\%$$

Be careful with δ. If too small, you always reject!

For this reason, there is an alternate procedure for 2-sided testing to arrive at a decision [redacted]

Retain H_0 at level α if $\theta_0 \in CR_{\theta, 1-\alpha}$ otherwise Reject H_0.

In our case $0.5 \in [.194, .932] \Rightarrow$ Retain H_0.

Done with hypothesis testing!!

You can get a pvalue out of this by taking the maximum α s.t. θ_0 is in the region CR_θ,1-α

DGP: $X_1 \dots X_n \overset{iid}{\sim} Bern(\theta)$, $f(\theta) = U(0,1)$.

Imagine $n=3$ $\vec{x} = (x_1, x_2, x_3)$ and do the inference one data point at a time using the posterior from the previous posterior as the prior:

$$f(\theta|x_1) = \frac{p(x_1|\theta) f(\theta)}{p(x_1)} \propto p(x_1|\theta) f(\theta) = \theta^{x_1}(1-\theta)^{1-x_1} \mathbb{1}_{\theta \in (0,1)} \propto beta(x_1+1, 1-x_1+1)$$

$$f(\theta|x_2) \propto p(x_2|\theta) f(\theta|x_1) \propto \theta^{x_2}(1-\theta)^{1-x_2} \theta^{x_1}(1-\theta)^{1-x_1} \mathbb{1}_{\theta \in (0,1)} \propto Beta\left(x_1+x_2+1,\ 2-(x_1+x_2)+1\right)$$

$$f(\theta|x_3) \propto p(x_3|\theta) f(\theta|x_2) \propto \theta^{x_3}(1-\theta)^{1-x_3} \theta^{x_1+x_2}(1-\theta)^{2-(x_1+x_2)} \mathbb{1}_{\theta \in (0,1)} \propto Beta\left(x_1+x_2+x_3+1,\ 3-(x_1+x_2+x_3)+1\right)$$

This process of iterative updating is not specific to this DGP with Laplace's prior, but holds always (Rev).

What else did we learn? beta prior $\rightarrow$ beta posterior.

Let's prove this in general:

DGP: $X_1 \dots X_n \overset{iid}{\sim} Bern(\theta)$, $f(\theta) = Beta(\alpha, \beta)$ $\overset{\text{if } \alpha=\beta=1}{=} U(0,1)$ special case

(α, β: the parameters on the prior are called "hyperparameters")

$$f(\theta|\vec{x}) \propto p(\vec{x}|\theta) f(\theta) \propto \left(\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\right)\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbb{1}_{\theta\in(0,1)}\right)$$

$$\propto Beta(\alpha + \textstyle\sum x_i,\ \beta + n - \sum x_i)$$

$$\Rightarrow \hat{\theta}^{MAP} = \frac{\alpha + \sum x_i - 1}{\alpha+\beta+n-2}, \quad \hat{\theta}^{MMSE} = \frac{\alpha+\sum x_i}{\alpha+\beta+n}, \quad \hat{\theta}^{median} = qbeta(0.5,\ \alpha+\textstyle\sum x_i,\ (\beta+n-\sum x_i)$$

when $\alpha + \sum x_i > 1$

We say the Beta distr is the "conjugate prior" for the iid Bernoulli DGP. Conjugacy means prior and posterior are same rv (but different parameter values updated by the data).

$$\text{Beta}(\alpha,\beta) \xrightarrow{\vec{x}} \text{beta}\left(\alpha + \underbrace{\textstyle\sum x_i}_{\#\text{successes}},\ \beta + n - \underbrace{\textstyle\sum x_i}_{\#\text{failures}}\right)$$

(# pseudosuccesses, # pseudofailures)

What is the interpretation of the values of the hyperparameters?

The prior's hyperparameters are like observing fake data, where $n_0 = \alpha+\beta$ # pseudotrials

If we employ Laplace's prior of indifference $\pi(\theta,1) = \text{Beta}(\alpha=1,\beta=1)$

$\Rightarrow n_0 = 2$ pseudotrials. This is weird. Further, consider

$$\hat{\theta}^{\text{MMSE}} = \frac{\alpha+\sum x_i}{\alpha+\beta+n} = \frac{\alpha}{\alpha+\beta+n}\cdot\frac{\alpha+\beta}{\alpha+\beta} + \frac{\sum x_i}{\alpha+\beta+n}\cdot\frac{n}{n}$$

$$= \frac{\alpha+\beta}{\alpha+\beta+n}\,\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\,\frac{\sum x_i}{n}$$

$$= \varrho\, \underbrace{E[\theta]}_{\text{prior guess}} + (1-\varrho)\,\underbrace{\hat{\theta}^{\text{MLE}}}_{\text{pure data estimate}}$$

Let $\varrho := \frac{\alpha+\beta}{\alpha+\beta+n}$, the shrinkage metric, measures strength of prior $\in[0,1]$ on the most popular point estimate.

e.g. $\alpha=\beta=1$, $n=3 \Rightarrow \varrho = \frac{2}{5}$

$\Rightarrow$ 40% weight on prior, 60% weight on data

$\lim_{n\to\infty} \hat{\theta}^{\text{MMSE}} = \hat{\theta}^{\text{MLE}}$ since $\lim_{n\to\infty} \varrho = 0$

$\hat{\theta}^{\text{MMSE}}$ is called a "shrinkage estimator" since it shrinks toward $E(\theta)$