X = iid Bernoulli

$P(\theta) = U(0,1) = Beta(1,1)$

$\Rightarrow P(\theta|x) \propto Beta(\sum x_i + 1,\ n - \sum x_i + 1)$

Lec 6 2/11/19 March 3rd

From last one

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$$ A mathematics def. This resolves to a #

$$Y \sim beta(\alpha,\beta) := \underbrace{\frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}}_{p(y):\text{ PDF of the Beta}}$$ Supp(Y) = [0,1], Param space: $\alpha,\beta > 0$

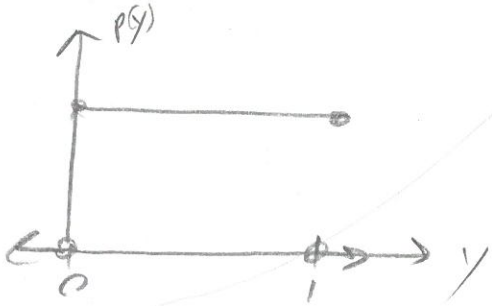
is a r.v.

no closed form

$E[Y] = \frac{\alpha}{\alpha+\beta}$, $Mode(Y) = \frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha,\beta > 1$, $Med(Y) = qbeta(0.5, \alpha, \beta)$

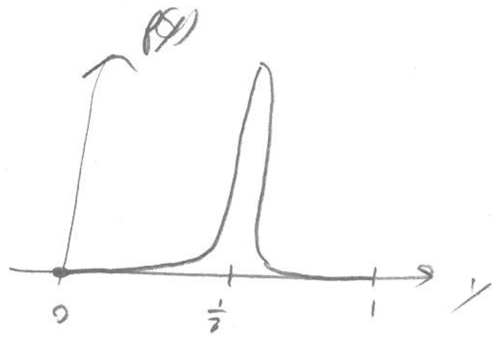
Common shapes

$\alpha = \beta = 1$



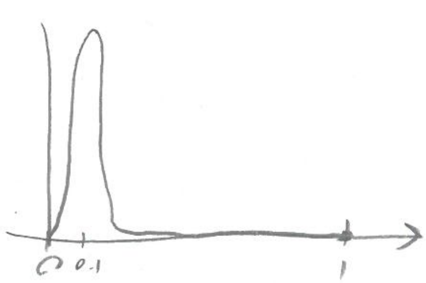
Beta(1,1) = U(0,1)

Std unif. special case of Beta

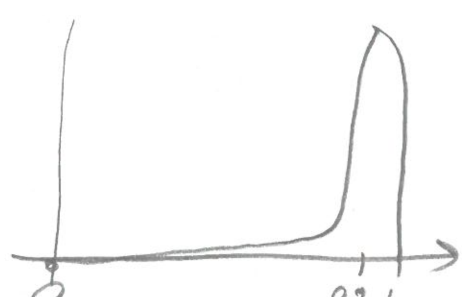


$\alpha = \beta = 100$

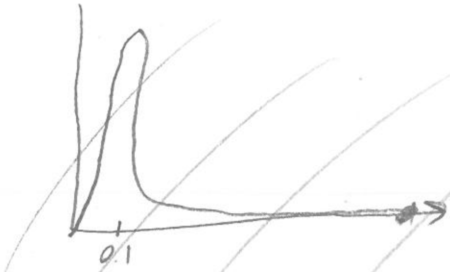
concentrates around $\frac{1}{2}$



$\alpha = 10, \beta = 90$ $E(Y) = 0.1$

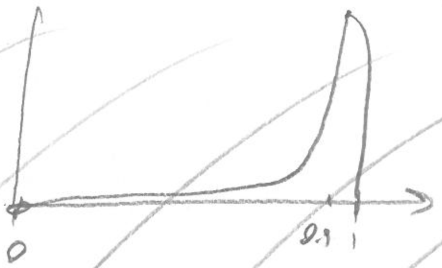


$\alpha = 90, \beta = 10$ $E(Y) = 0.9$



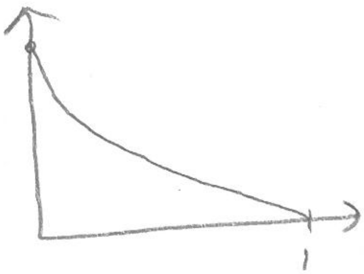
$\alpha = 10, \beta = 90$

concentrate about 0.1

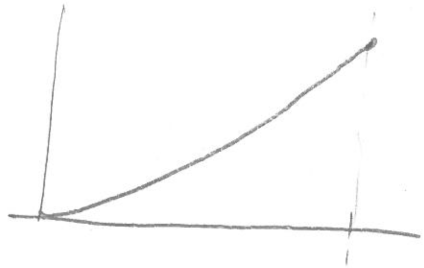


$\alpha = 90, \beta = 10$

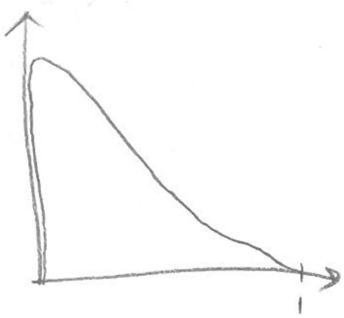
concentrate about 0.9



$\alpha = 1, \beta = 4$



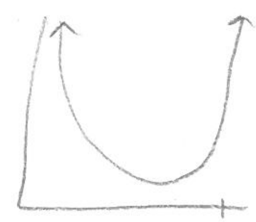
$\alpha = 4, \beta = 1$



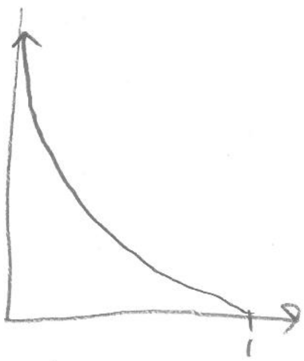
$\alpha = 1.001, \beta = 4$

$\alpha = 4, \beta = 1.001$

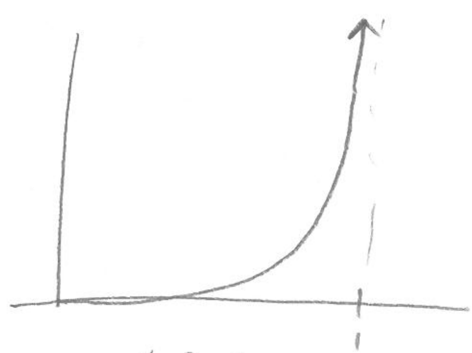
Arcsin distr.



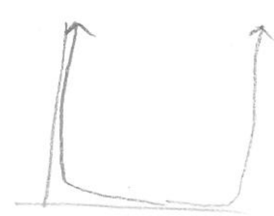
$\alpha = \beta = \frac{1}{2}$



$\alpha = 0.999, \beta = 3$



$\alpha = 3, \beta = 0.999$



$\alpha = \beta = \frac{1}{100}$

Recall problem #1 with frequentist inference.

Let's return to $X = \langle 0,0,0 \rangle$ $\hat{\theta}_{MLE} = \bar{x} = 0$ why not a good idea?

Let's see what happens with Bayesian pt. est.

$$P(\theta|x) = \text{Beta}\left(\sum x_i + 1,\ n - \sum x_i + 1\right) = \text{Beta}(1, 4)$$

$$\hat{\theta}_{MMSE} = E[\theta|x] = \frac{1}{1+4} = 0.2$$ ← Very reasonable.

$$\hat{\theta}_{MAP} = \text{Mode}[\theta|x] \overset{= \hat{\theta}_{MLE}}{=} 0 \quad \text{(see picture)}$$

$$\hat{\theta}_{MMAE} = \text{Med}[\theta|x] = \text{qbeta}(0.5, 1, 4) = .1591036\ldots$$

Why? Leaves door open to the event $x=1$.

In the book... talked about insurance policies for midair crashes just when planes started to fly commercially ~1930's. Same case?

Recall

$$P(\theta|x_1) = \frac{P(x_1|\theta)\,P(\theta)}{P(x_1)} = \text{Beta}(1, 2)$$

$$P(\theta|x_2) = \frac{P(x_2|\theta)\,P(\theta|x_1)}{P(x_2)} = \frac{P(x_1, x_2|\theta)\,P(\theta)}{P(x_1, x_2)} = \text{Beta}(1,3)$$

$$P(\theta|x_3) = \frac{P(x_3|\theta)\,P(\theta|x_2)}{P(x_3)} = \frac{P(x_1, x_2, x_3|\theta)\,P(\theta)}{P(x_1, x_2, x_3)} = \text{Beta}(1,4)$$

It seems that a beta prior yields a beta posterior for $\mathcal{F}$ = iid Bernoulli. Let's generalize this...

let $P(\theta) = \text{Beta}(\alpha, \beta)$, $X_1, \ldots, X_n$ draw

$$P(\theta \mid x) = \frac{P(x|\theta)P(\theta)}{P(x)} = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \cancel{\frac{1}{B(\alpha,\beta)}} \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \cancel{\frac{1}{B(\alpha,\beta)}} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta} = \frac{\theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1}}{\int_0^1 \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1} d\theta}$$

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

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

$$\overbrace{\text{Beta}(\alpha,\beta)}^{P(\theta)} \xrightarrow{\;x\;} \overbrace{\text{Beta}\left(\sum x_i+\alpha, n-\sum x_i+\beta\right)}^{P(\theta|x)}$$

"Beta" is the "conjugate prior" for the "iid Bernoulli likelihood model".

Conjugacy: prior and posterior share the same r.v.

α, β are params of the prior distr. They are called hyperparameters. Who specifies them?

Back to Math 241 ...

If $X_1, \ldots, X_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta) \Rightarrow X = \sum_{i=1}^n X_i \sim \text{Bin}(n, \theta)$

$$= \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

$\mathcal{F}$ = Binom w/ fixed n, $P(\theta) = \text{Beta}(\alpha, \beta)$

let "x" represent the sum

$$P(\theta \mid x) = \frac{P(x|\theta)P(\theta)}{P(x)} = \frac{\cancel{\binom{n}{x}}\theta^x(1-\theta)^{n-x} \cancel{\frac{1}{B(\alpha,\beta)}} \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \cancel{\binom{n}{x}}\theta^x(1-\theta)^{n-x} \cancel{\frac{1}{B(\alpha,\beta)}} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta}$$

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

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

"Beta is conjugate for the binomial likelihood"

$$\overbrace{\text{Beta}(\alpha,\beta)}^{p(\theta)} \xrightarrow{\;x\;} \overbrace{\text{Beta}(\underbrace{x}_{\text{\# of data successes}}+\underbrace{\alpha}_{\text{\# of prior successes}},\ \underbrace{n-x}_{\text{\# of data failures}}+\underbrace{\beta}_{\text{\# of prior failures}})}^{p(\theta|x)}$$

\# of prior successes (pseudosuccesses) (pseudocounts)

\# of prior failures (pseudofailures)

Units?

Hyperparameters for conjugate distributions have a special interpretation; they represent pseudodata

Bayesian Pt Est's

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

$$\text{Mode}[\theta|x] = \frac{x+\alpha-1}{n+\alpha+\beta-2} \quad \text{if } \alpha+x>1,\ \beta+n-x>1$$

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

$U(0,1) = \text{Beta}(1,1)$ We thought this was the prior for the principle of indifference. But how many pseudosuccesses does it represent? 1

" " " pseudofailures does it represent? 1

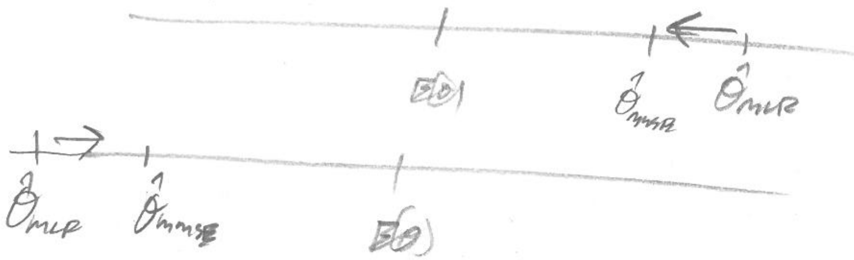
How much pseudodata does it represent? $n_0 = 2$, $E[\theta] = \frac{1}{2}$

It definitely contains information! The principle of indifference is not so indifferent.

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

$$= \underbrace{\frac{n}{n+\alpha+\beta}}_{1-\rho}\cdot\underbrace{\frac{x}{n}}_{\hat{\theta}_{MLE}} + \underbrace{\frac{\alpha+\beta}{n+\alpha+\beta}}_{\rho}\cdot\underbrace{\frac{\alpha}{\alpha+\beta}}_{E[\theta]} = (1-\rho)\,\hat{\theta}_{MLE} + \rho\, E[\theta]$$

$\hat{\theta}_{MMSE}$ is known as a "shrinkage estimator". It takes the $\hat{\theta}_{MLE}$ and shrinks towards $E[\theta]$, the prior idea.



"shrink down"

"shrink up"

"ρ" is the shrinkage proportion. How does it behave?

If n large $\rho \simeq 0$. Makes sense? $\hat{\theta}_{MMSE} \to \hat{\theta}_{MLE}$ if n large.

Data drowns out prior

If n small and α, β small, $\hat{\theta}_{MLE}$ still is strongly represented

If n small and α, β large, the estimate is shrunk hard towards $E[\theta]$.

α, β "small" means uninformative prior e.g. $U(0,1)$ → doesn't affect point estimate

α, β "large" means informative prior → affects point estimate.

Why would you want an "informative prior"? We will see next class.