Likelihood, Prior to Posterior probability, Posterior Distributions

What will be discussed…

- Bayesian inference :

How does the Bayesian theorem work to obtain posterior information about unknown parameters?

  - Example for discrete parameters
  - Beta-Binomial Bayesian model

- Likelihood function : how do we form a likelihood given observed data?

  - kernel and normalizing constant

  - In Bayesian statistics spotting kernels of distributions can be very useful in deriving posterior distributions.

(Recap) Bayesian Method for Inference

1. Prior : $[\theta]$, $f(\theta)$

Specify a prior distribution which expresses our knowledge about θ prior to observing the data.

2. Likelihood : $[x|\theta]$, $f(x|\theta)$

Model a set of observations with a probability distribution (expressed in the form of the likelihood function) with unknown parameter(s)

3. Posterior : $[\theta|x]$, $f(\theta|x) = \dfrac{f(x|\theta)f(\theta)}{f(x)}$

Apply Bayes' theorem to derive posterior distribution which expresses all that is known about θ after observing the data.

4. Inference: Derive inference from posterior distribution.

e.g. point/interval estimates, probabilities of specified hypotheses.

Bayes' Theorem in Parametric Distributions

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{1}{f(x)}f(x|\theta)f(\theta)$$

Posterior of θ = normalizing constant · likelihood · prior

$\Longrightarrow$ Posterior $\propto$ Likelihood · Prior

$$f(\theta|x) \propto f(x|\theta)f(\theta)$$

- $f(x)$ is a constant with respect to θ:

$$f(x) = \int f(x,\theta)d\theta = \int f(x|\theta)f(\theta)d\theta$$

$$\Longrightarrow \quad f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \quad \text{for continuous } \theta$$

Review: joint, marginal, conditional density (Aside)

Let X and Y be random variables with the joint density $f_{XY}(x,y)$.

- The marginal density of X is $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy$

- The conditional density of Y given $X = x$: $f(y|x) = \dfrac{f_{XY}(x,y)}{f_X(x)} = \dfrac{f(x|y) f_Y(y)}{f_X(x)}$

- When X and Y are independent,
  - $f_{XY}(x,y) = f_X(x)\, f_Y(y)$
  - $f(y|x) = f_Y(y)$
  - $f(x|y) = f_X(x)$

- When X and Y are (conditionally) independent given Z,

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

The Likelihood Function

- Suppose that $X_1,\ldots,X_n$ are from a distribution with $f(x: \theta)$,

a probability mass function (pmf) for a discrete random variable (rv) X, or a probability density function (pdf) for a continuous X.

- Def: Given that $\boldsymbol{X} = \boldsymbol{x}$ (i.e. $X_1 = x_1,\ldots,X_n = x_n$), the function of θ defined by

$$L(\theta) \equiv L(\theta : \boldsymbol{x}) = k\, f(\boldsymbol{x} : \theta)$$

is called the likelihood function, where $k > 0$ and k does not depend on θ.

- The likelihood function $L(\theta : \boldsymbol{x})$ is formed from the joint pdf or pmf of X, but is viewed as a function of θ with data $X_1 = x_1,\ldots,X_n = x_n$ held fixed.

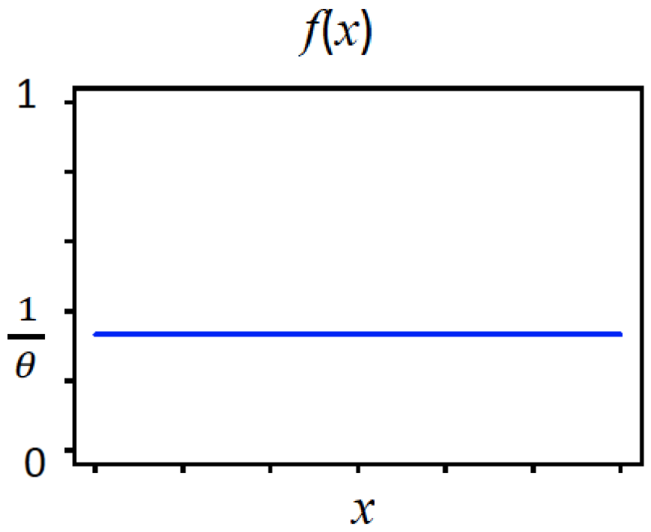
- The pmf or pdf $f(\boldsymbol{x}: \theta)$ is a model that describes the random behavior of X when θ is fixed.

Example 1

- $X \sim \text{Unif}(0, \theta)$

i) The pdf of X is $f(x: \theta) = \frac{1}{\theta}$ for $0 < x < \theta$

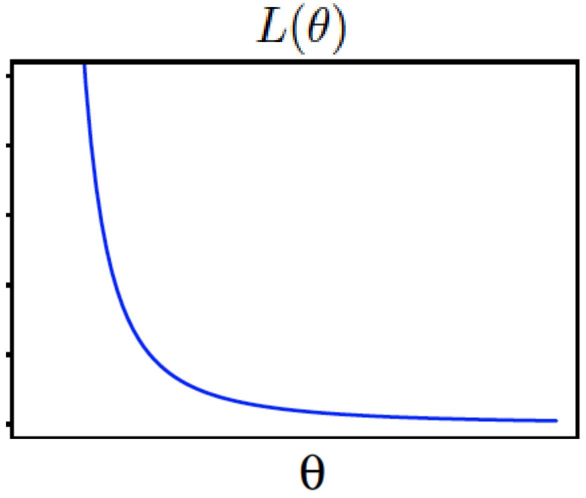
(θ is fixed)



ii) The likelihood function is $L(\theta : x) = \frac{1}{\theta}$

for $x < \theta$

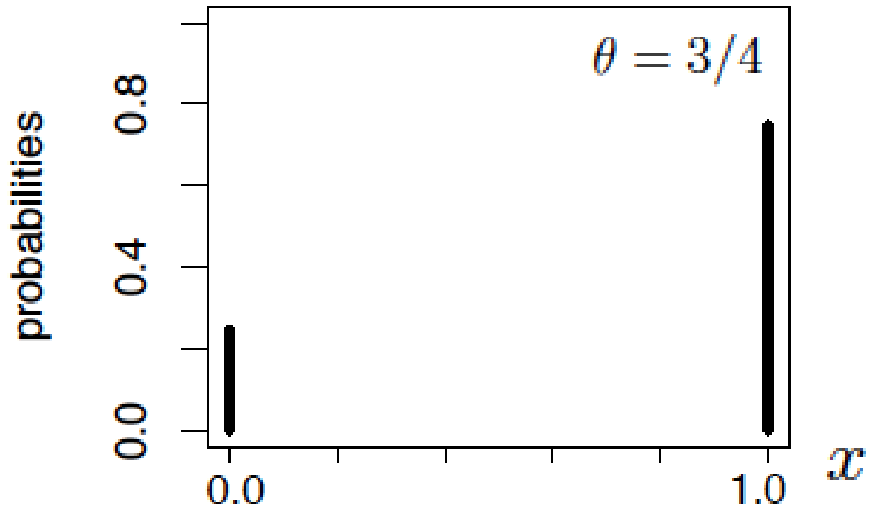
(x is fixed)



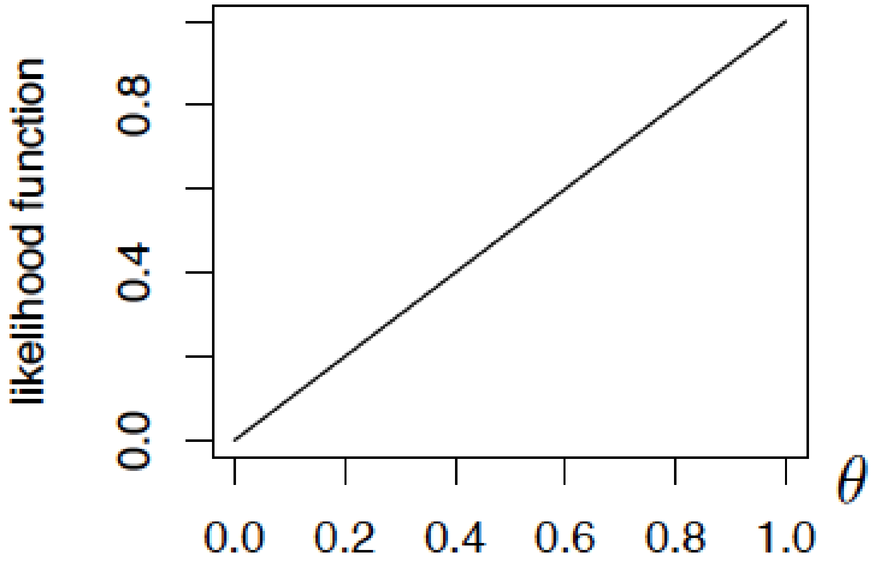
E.g. Flip a coin : $X = \begin{cases} 0 & \text{if tail} \\ 1 & \text{if head} \end{cases}$

- Let θ be the probability of head: $\begin{cases} P(X = 0|\theta) = 1 - \theta & \text{if tail} \\ P(X = 1|\theta) = \theta & \text{if head} \end{cases}$

$$\Longrightarrow \quad P(X = x|\theta) = \theta^x(1-\theta)^{1-x} \quad \text{where } x = 0 \text{ or } 1$$



Hold θ constant / vary x



Fix x=1

Jargon

- $X_1, \ldots, X_n$ are are independent and identically distributed (*i.i.d.*):

Then, $f_{\boldsymbol{X}}(x_1, \ldots, x_n) = \underbrace{f_1(x_1) f_2(x_2) \cdots f_n(x_n)}_{\text{independent}} = \underbrace{f(x_1) f(x_2) \cdots f(x_n)}_{\text{identical}}$

$$\Leftrightarrow \quad f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_i(x_i\colon \theta) = \prod_{i=1}^{n} f(x_i\colon \theta)$$

$$\Leftrightarrow \quad L(\theta\colon \boldsymbol{x}) = \prod_{i=1}^{n} L(\theta\colon x_i) = \prod_{i=1}^{n} f(x_i\colon \theta)$$

- E.g. $X_1,\ldots,X_n \sim i.i.d. Uniform(0, \theta)$: $f(x\colon \theta) = \frac{1}{\theta}$ for $0 < x < \theta$

$$\Rightarrow \quad L(\theta\colon \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i\colon \theta) = \left[\frac{1}{\theta}\right]^n$$

iii) Suppose that $X_1, \ldots, X_6 \sim\ i.i.d.$ Poisson(λ) i.e. $f\,(x_i|\lambda) = \dfrac{e^{-\lambda}\,\lambda^{x_i}}{x_i\,!}$

Then, $L(\lambda\colon (x_1\,, \ldots, x_6)\,) \equiv\ f\,(\boldsymbol{x}\colon\ \lambda)$

Or, $L(\lambda\colon \boldsymbol{x})$ = $\prod_{i=1}^{6} f\,(x_i\colon\lambda)$ =

iv) Suppose that $X_1, \ldots, X_n \sim\ i.i.d.$ Gamma(α, 1/β)

$$L(\alpha, \beta\colon \boldsymbol{x}) = \prod f\,(x_i) = \prod \left[\frac{\beta^{\alpha}}{\Gamma(\alpha)}\, x_i^{\alpha-1} \exp(-\beta x_i)\right]$$

$$=$$

iii) Suppose that $X_1,\ldots, X_6 \sim i.i.d.$ Poisson(λ)

Then, $L(\lambda: (x_1, \ldots, x_6)) \equiv f(\boldsymbol{x}: \lambda) = \dfrac{\lambda^{x_1} e^{-\lambda}}{x_1!} \times \cdots \times \dfrac{\lambda^{x_6} e^{-\lambda}}{x_6!} = \dfrac{\lambda^{\sum_{i=1}^{6} x_i} e^{-6\lambda}}{x_1! \cdots x_6!}$

$$\text{Or, } L(\lambda: \boldsymbol{x}) = \prod_{i=1}^{6} f(x_i:\lambda) = \prod_{i=1}^{6} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{6} x_i} e^{-6\lambda}}{\prod x_i!}$$

iv) Suppose that $X_1,\ldots, X_n \sim i.i.d.$ Gamma(α, 1/β)

$$L(\alpha, \beta: \boldsymbol{x}) = \prod f(x_i) = \prod \left[\frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \exp(-\beta x_i)\right]$$

$$= \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right]^n \left(\prod x_i^{\alpha-1}\right) \exp\left(-\beta \sum_{i=1}^{n} x_i\right)$$

Inference with Likelihood function

- The likelihood function $L(\theta: \boldsymbol{x})$ is a function of θ that shows how "likely" various parameter values of θ may have produced the data $\boldsymbol{x}$ that were observed.

- In classical (frequentist) statistics, the specific value of θ that maximizes $L(\theta: \boldsymbol{x})$ is the maximum likelihood estimator (MLE) of θ.

  Here, we ask "what value of θ makes the data most likely to occur?"

- In a Bayesian context, we are interested in:

  "what value of θ is most likely given the data?"

- In a classical analysis this question makes no sense, since all the randomness within $L(\theta|\, \boldsymbol{x})$ is attached to X, not to θ.

Example 2 Bayesian method for Discrete parameters

- Suppose that there are three states of nature A_1, A_2, A_3 and two possible data D_1, D_2:

	$P(D\|A)$		
	D_1	D_2	Prior
A_1	0.0	1.0	0.3
A_2	0.7	0.3	0.5
A_3	0.2	0.8	0.2

- What happens to our belief about A_1, A_2, A_3 if we observe D_2? (if we observe D_1?)

Posterior Probabilities with D_2

	Likelihood	Prior	Lkhd x prior (joint)	Posterior
A_1	1.0	0.3	0.3	
A_2	0.3	0.5		
A_3	0.8	0.2		
		1	P(D_2) = 0.61	1

- P($A_1|D_2$) = $\dfrac{\text{P}[D_2|A_1]\text{P}(A_1)}{\text{P}[D_2]}$

Posterior Probabilities with D_2

	Likelihood	Prior	Lkhd x prior (joint)	Posterior
A_1	1.0	0.3	0.3	0.3/0.61 ≈ 0.4918
A_2	0.3	0.5	0.15	0.15/0.61 ≈ 0.2459
A_3	0.8	0.2	0.16	0.16/0.61≈ 0.2623
		1	P(D_2) = 0.61	1

- $$\mathrm{P}(A_1|D_2) = \frac{\mathrm{P}[D_2|A_1]\mathrm{P}(A_1)}{\mathrm{P}[D_2]} = \frac{\mathrm{P}[D_2|A_1]\mathrm{P}(A_1)}{\mathrm{P}[D_2, A_1]+\mathrm{P}[D_2, A_2]+\mathrm{P}[D_2, A_3]}$$

$$= \frac{\mathrm{P}[D_2|A_1]\mathrm{P}(A_1)}{\mathrm{P}[D_2|A_1]\mathrm{P}(A_1)+\mathrm{P}[D_2|A_2]\mathrm{P}(A_2)+\mathrm{P}[D_2|A_3]\mathrm{P}(A_3)}$$

$$= \frac{0.3}{0.3+0.3\mathrm{x}0.5+0.8\mathrm{x}0.2} = 0.4918$$

Posterior Probabilities with D_1

	Likelihood	Prior	Lkhd x prior (joint)	Posterior
A_1	0.0	0.3	0	0
A_2	0.7	0.5	0.35	0.35/0.39 ≈ 0.8974
A_3	0.2	0.2	0.04	0.04/0.39≈ 0.1026
		1	P(D_1) = 0.39	1

- $\mathrm{P}(A_2|D_1) = \dfrac{\mathrm{P}[D_1|A_2]\mathrm{P}(A_2)}{\mathrm{P}[D_1\,]} = \dfrac{\mathrm{P}[D_1|A_2]\mathrm{P}(A_2)}{\mathrm{P}[D_1\,,A_1]+\mathrm{P}[D_1\,,A_2]+\mathrm{P}[D_1\,,A_3]}$

$$= \frac{\mathrm{P}[D_1|A_2]\mathrm{P}(A_2)}{\mathrm{P}[D_1|A_1]\mathrm{P}(A_1)+\mathrm{P}[D_1|A_2]\mathrm{P}(A_2)+\mathrm{P}[D_1|A_3]\mathrm{P}(A_3)}$$

$$= \frac{0.7\text{x}0.5}{0+\ 0.7\text{x}0.5\ +0.2\text{x}0.2} = 0.8974$$

Example 3

- A black male mouse is mated with a female black mouse whose mother had a brown coat.

- B and b are alleles of the gene for coat color. The gene for black fur is given the letter B and the gene for brown fur is given the letter b where B is the dominant allele to b. The mouse is brown only if it is homozygous bb.

- The male and female have a litter with 5 pups that are all black. We want to determine the male's genotype.

- The prior information suggests that P(BB) = 1/3 and P(Bb) = 2/3.

Q. What is the posterior probability that the male's genotype is BB?

- Black female's mother is brown (Mother: bb) $\Longrightarrow$ Black female must be Bb.

- Litter of 5 pups are all black:

Male	Female		Pup	Prob. of a black pup
$\left.\begin{matrix}BB\\Bb\end{matrix}\right\}$	Bb	$\Longrightarrow$	$\left.\begin{matrix}BB \text{ or } Bb\\BB, Bb, bB, bb\end{matrix}\right\}$	$\left.\begin{matrix}1\\\frac{3}{4}\end{matrix}\right\}$

- Lkhd: P(pup 1 black,…, pup 5 black) = P(pup 1 is black)x⋯xP(pup 5 is black)

Male	Likelihood	Prior	Lkhd x prior	posterior
BB	1^5	$\frac{1}{3}$		
Bb	$(\frac{3}{4})^5$	$\frac{2}{3}$		
		sum to 1		sum to 1

- P(male is BB| 5 pups are black) =?

- Black female's mother is brown (Mother: bb) $\Longrightarrow$ Black female must be Bb.

- Litter of 5 pups are all black:

Male	Female		Pup	Prob. of a black pup
BB	Bb	$\Longrightarrow$	BB or Bb	1
Bb			BB,Bb,bB, bb	$\frac{3}{4}$

- Lkhd: P(pup 1 black,…, pup 5 black) = P(pup 1 is black)x⋯xP(pup 5 is black)

Male	Likelihood	Prior	Lkhd x prior	posterior
BB	1^5	$\frac{1}{3}$	0.333	0.333/0.491 ≈ 0.678
Bb	$(\frac{3}{4})^5$	$\frac{2}{3}$	0.158	0.158/0.491 ≈ 0.322
		sum to 1	0.491	sum to 1

- P(male is BB| 5 pups are black) ≈ 0.678 (updated from 0.333)

Kernel & Normalizing Constant

- For a random variable X with density (or mass) function $f_X(x)$:

(1) $f_X(x) \geq 0$ (must be nonnegative) for each value of random variable (rv) X

(2) - $\Sigma f_X(x) = 1$ for a discrete rv.

- $\int f_X(x)dx = 1$ for continuous rv.

- If $f(\mathbf{x}|\theta)$ can be expressed in the form $cq(\mathbf{x}|\theta)$ where c is a constant, not depending upon $\boldsymbol{x}$, then any such $q(\mathbf{x}|\theta)$ is a kernel of the density $f(\mathbf{x}|\theta)$.

The constant c is called a normalizing constant with the fact

$$\int f(\mathbf{x}|\theta)\, dx = \int cq(\mathbf{x}|\theta)\, dx = 1 \quad \Longrightarrow \quad \int q(\mathbf{x}|\theta)dx = \frac{1}{c}$$

For the discrete case, the integral is replaced by a sum.

- In Bayesian statistics spotting kernels of distributions can be very useful in computing/finding posterior distributions.

i) $Y_i \mid \lambda \sim \text{Poiss}(\lambda)$

$$f(y_i \mid \lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \qquad\qquad \sum_{y_i=0}^{\infty} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = 1 \quad \Longrightarrow ?$$

ii) $\theta \sim \text{Beta}(\alpha, \beta) \qquad 0 < \theta < 1$

$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

- In Bayesian statistics spotting kernels of distributions can be very useful in computing/finding posterior distributions.

i) $Y_i | \lambda \sim \text{Poiss}(\lambda)$

$$\sum_{y_i=0}^{\infty} f(y_i|\lambda) = \sum_{y_i=0}^{\infty} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = 1 \implies \underbrace{e^{-\lambda}}_{\text{normalizing constant (n.c.)}} \sum_{y_i=0}^{\infty} \frac{\overbrace{\lambda^{y_i}}^{\text{kernel}}}{y_i!} = 1$$

ii) $\theta \sim \text{Beta}(\alpha, \beta)$ $\quad 0 < \theta < 1$

$$\text{P}(\theta) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\text{normalizing constant}} \underbrace{\theta^{\alpha-1} (1-\theta)^{\beta-1}}_{\text{kernel}} \implies \int_0^1 \theta^{\alpha-1} (1-\theta)^{\beta-1} \, d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

i.e. the integral of a kernel = 1/n.c.

List of some probability distributions (p12 in notes)

- $Y_i|(\alpha,\beta) \sim$ Gamma(α,β) distribution

$$f(y_i|\alpha,\beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)} y_i^{\alpha-1}\exp(-y_i/\beta) \quad y_i > 0, \quad \alpha > 0, \beta > 0$$

$E[Y_i|(\alpha,\beta)] = \alpha\beta, \quad Var[Y_i|(\alpha,\beta)] = \alpha\beta^2.$

- $Y_i|(\alpha,\beta) \sim$ Inverse Gamma(α,β) distribution

$$f(y_i|\alpha,\beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)} y_i^{-(\alpha+1)}\exp(1/(-y_i\beta)) \quad y_i > 0, \quad \alpha > 0, \beta > 0$$

$E[Y_i|(\alpha,\beta)] = \frac{1}{(\alpha-1)\beta}, \quad Var[Y_i|(\alpha,\beta)] = \frac{1}{(\alpha-1)^2(\alpha-2)\beta^2}.$

- $Y_i|(\mu,\sigma^2) \sim$ Normal (μ,σ^2) distribution

$$f(y_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right), \quad -\infty < y_i < \infty \quad \mu > 0, \sigma^2 > 0.$$

$E[Y_i|(\mu,\sigma^2)] = \mu, \quad Var[Y_i|(\mu,\sigma^2)] = \sigma^2.$

- $Y_i|\lambda \sim$ Poisson (λ) distribution

$$f(y_i|\lambda) = \lambda^{y_i} e^{-\lambda}/y_i! \quad y_i = 0, 1, 2, ...$$

$$E[Y_i|\lambda] = \lambda, \quad Var[Y_i|\lambda] = \lambda.$$

- $Y_i|p \sim$ Binomial (n, p) distribution

$$f(y_i|n, p) = \binom{n}{y_i} p^{y_i}(1-p)^{n-y_i}, \quad y_i = 0, 1, 2, ...$$

$$E[Y_i|p] = np, \quad Var[Y_i|p] = np(1-p).$$

- $Y_i|(\alpha, \beta) \sim$ Beta (α,β) distribution

$$f(y_i|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y_i^{\alpha-1}(1-y_i)^{\beta-1}, \quad 0 < y_i < 1 \quad \alpha > 0, \beta > 0.$$

$$E[Y_i|(\alpha, \beta)] = \tfrac{\alpha}{\alpha+\beta}, \quad Var[Y_i|(\alpha, \beta)] = \tfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Examples

- Write down the probability density function and find a correseponding kernel. Refer to the list of probability distributions (in p11 of course notes)

1) $\phi \sim \text{Gamma}(b + x, \frac{1}{2d})$

2) $\lambda \sim \text{Normal}(\frac{1}{a}, \frac{1}{b^2})$

3) The pdf of θ is

$$\frac{1}{\left(\frac{1}{\beta}\right)^{\alpha+y-1}\Gamma(\alpha + y - 1)} \theta^{-(\alpha+y-1+1)} \exp\left(\frac{-1}{\theta\left(\frac{1}{\beta}\right)}\right)$$

What distribution does θ follow? Find the kernel of the density.

1) $$f(\phi) = \frac{1}{\Gamma(b+x)\left(\frac{1}{2d}\right)^{b+x}} \phi^{b+x-1} \exp\left(\frac{-\phi}{\left(\frac{1}{2d}\right)}\right)$$

$$= \frac{(2d)^{b+x}}{\Gamma(b+x)} \phi^{b+x-1} \exp(-2d\,\phi) = \text{normalizing constant x kernel}$$

2) $$f(\lambda) = \frac{1}{\sqrt{2\pi\left(\frac{1}{b^2}\right)}} \exp\left(\frac{-\left(\lambda - \frac{1}{a}\right)^2}{2\left(\frac{1}{b^2}\right)}\right)$$

$$= \frac{b}{\sqrt{2\pi}} \exp\left(\frac{-b^2\left(\lambda - \frac{1}{a}\right)^2}{2}\right) = \text{n.c. x kernel}$$

3) $\theta \sim$ Inverse Gamma $(\alpha + y - 1, \frac{1}{\beta})$

Example : Binomial-Beta

- Suppose that $X|\theta \sim \text{Binom}(n, \theta)$: $f(x|\theta) = \binom{n}{x}\theta^x (1-\theta)^{n-x}$

- Since the parameter θ is restricted to be between 0 and 1, we should choose a prior distribution with support on [0, 1].

  - Can specify a prior distribution for θ : $\theta \sim \text{Beta}(\alpha, \beta)$ for $\alpha, \beta > 0$ known

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \text{for } 0 \leq \theta \leq 1$$

where α and β are the **hyperparameters** of this prior model, ideally reflecting our prior beliefs about θ.

- Using Bayes' theorem, the posterior is

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$$

Combine Prior & Likelihood

1. $f(x|\theta)f(\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$=$$

2. $\int f(x|\theta)f(\theta)d\theta = \int_0^1 \binom{n}{x}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\,d\theta$

$$=$$

Since $\int_0^1 f(\theta|x)d\theta = 1$, and $\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$ is a kernel of ?

$$\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\,d\theta = \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)} \quad : \frac{1}{normalizing\ const}$$

Combine Prior & Likelihood

1. $f(x|\theta)f(\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

2. $\displaystyle\int f(x|\theta)f(\theta)d\theta = \int_0^1 \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\,d\theta$

$$= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\,d\theta$$

Since $\int_0^1 f(\theta|x)d\theta=1$, and $\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$ is a kernel of Gamma dist.,

$$\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\,d\theta = \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)} \quad : \frac{1}{normalizing\ const}$$

Derive the Posterior Distribution

3. $f(\theta|x) = \dfrac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$

$$f(\theta|x) = \frac{\binom{n}{x}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\binom{n}{x}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\dfrac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)}}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$: density of Beta(x+α, n−x+β)

Thus, the posterior distribution : $\theta|x \sim$ Beta(x+α, n−x+β)

Short-cut to derive a posterior dist.

- $f(\theta|x) \propto f(x|\theta)f(\theta)$: posterior ∝ lkhd x prior

- Lkhd x prior: $f(x|\theta)f(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

Ignoring constants (in θ),

$$f(x|\theta)f(\theta) \propto \theta^x(1-\theta)^{n-x}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

: This is a kernel of Beta(x+α, $n-x$+β)

Then, the posterior distribution is $\theta|x \sim$ Beta(x+α, $n-x$+β)

Recap – to derive the posterior dist.

- $f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$

- The denominator $f(x)$ is just a normalizing constant and we don't actually have to calculate it (except posterior probabilities for discrete cases).

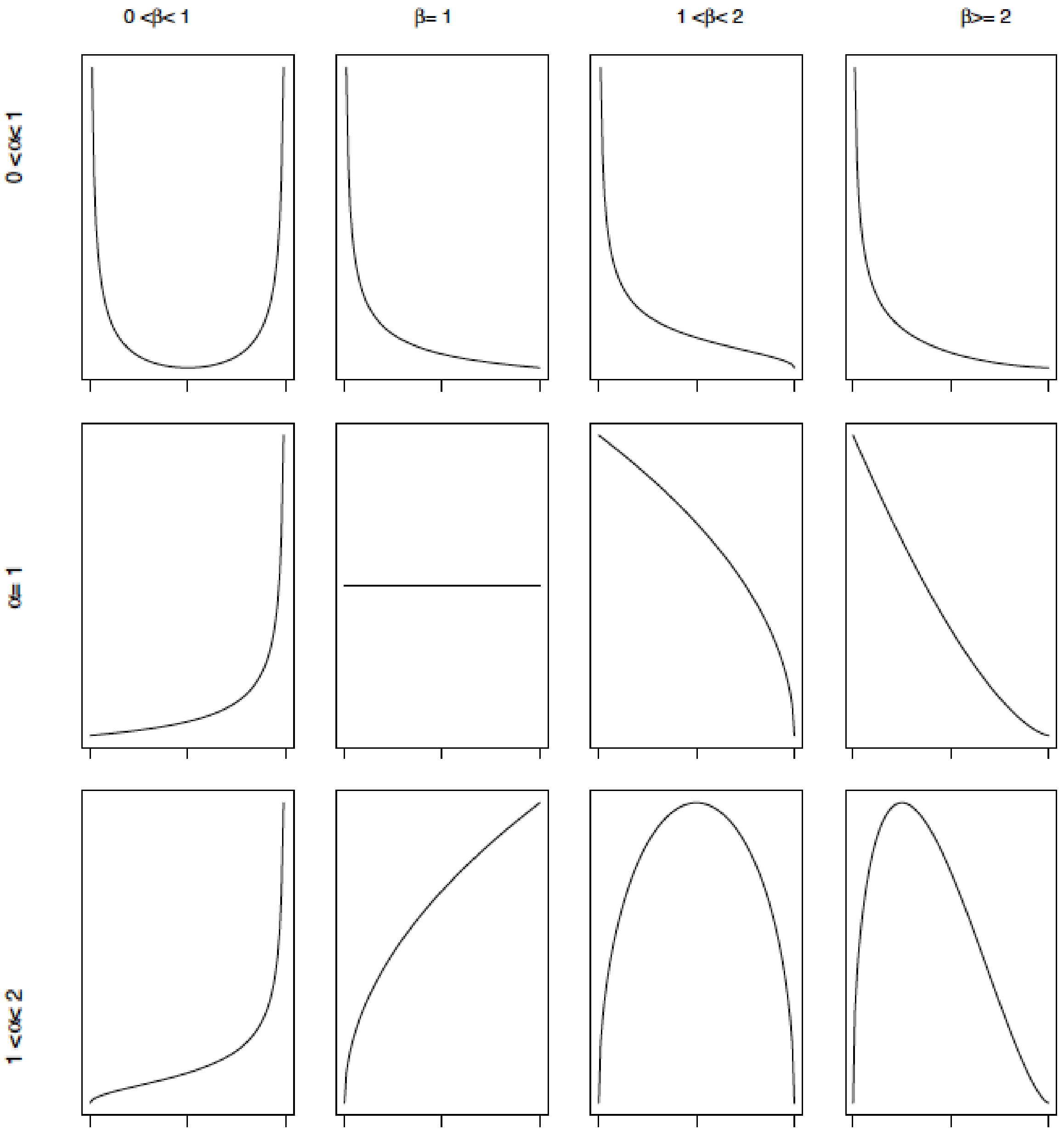
- We can use the fact that the posterior is proportional to the prior times the likelihood, i.e.

$$f(\theta|x) \propto f(x|\theta)f(\theta)$$: posterior ∝ lkhd x prior

- Notice that we can ignore all of the normalizing constants in the likelihood and the prior.

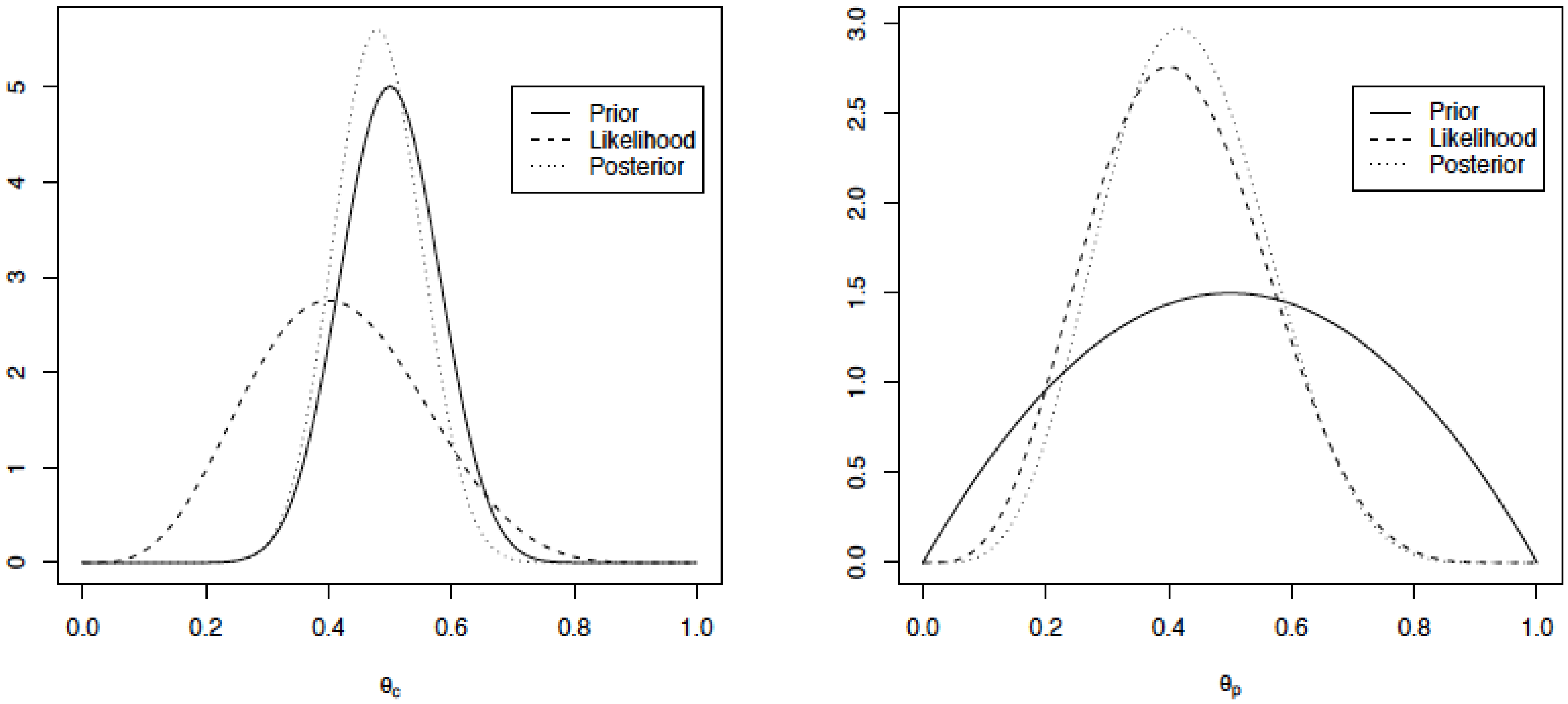
- This leaves us with only the kernel of the posterior distribution. This kernel leads us to identify the posterior distribution we want to find.

- Plots of the Beta pdf for various values of α and β can help inform the prior specification



Plots of prior, likelihood & posterior

- Eg. Observe 4 heads out of 10 tosses.



	Prior	*Likelihood*	*Posterior*
θ	$Beta(20, 20)$	$x = 4, n = 10$	$Beta(24, 26)$
θ	$Beta(2, 2)$	$x = 4, n = 10$	$Beta(6, 8)$