Prior and Posterior

PUBH 8442: Bayes Decision Theory and Data Analysis

Eric F. Lock
UMN Division of Biostatistics, SPH
elock@umn.edu

01/25/2021

Likelihood

- Assume a *sampling model* for data $\mathbf{y} = (y_1, \ldots, y_n)$
- Specified by parameters θ, which may be unknown
- Often represented as a probability density $p(\mathbf{y} \mid \theta)$
- E.g., Gaussian model specified by $\theta = (\mu, \sigma^2)$:

$$p(y_1 \mid \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_1 - \mu)^2}{2\sigma^2}\right\}$$

And under independent $y_i' s$:

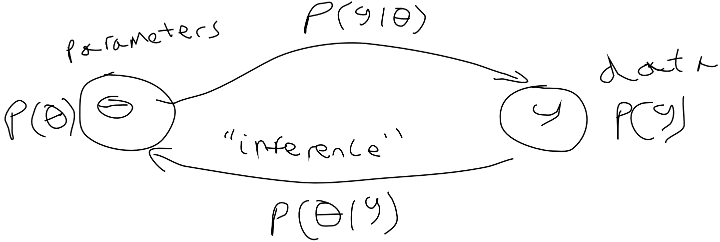
$$p(\mathbf{y} \mid \theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right\}$$

Likelihood

- The probability density $p(\mathbf{y} \mid \theta)$ is often called the *likelihood*
  - Sometimes with the notation $L(\theta; \mathbf{y})$
- Use this notation when making inferences about θ
- Can choose θ to maximize likelihood:

$$\hat{\theta} = \text{argmax}_\theta \, L(\theta; \mathbf{y})$$

  - i.e., estimate θ to maximize density of observed data
  - Called *maximum likelihood estimation* (MLE)
  - Has been criticized for overfitting



P(y|θ)
parameters
P(θ)
θ
data
y
P(y)
"inference"
P(θ|y)

Prior and posterior

- MLE approach assumes θ is fixed (though unknown)
- Alternatively, treat θ as a random variable
- Give θ a probability density $p(\theta)$
  - Potentially specified by *hyperparameters* η: $p(\theta \mid \eta)$.
- The *marginal* density of $\mathbf{y}$ is

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \theta) p(\theta) \, d\theta.$$

  - "averaging" over θ

Prior and posterior

- Bayes' rule for continuous random variables:

$$p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y}, \theta)}{p(\mathbf{y})}$$
$$= \frac{p(\mathbf{y} \mid \theta)p(\theta)}{\int p(\mathbf{y} \mid \theta)p(\theta)\, d\theta}.$$

- $p(\theta)$ is the prior, $p(\theta \mid \mathbf{y})$ the posterior distribution for θ
- $\int p(\mathbf{y} \mid \theta)p(\theta)\, d\theta$ is the normalizing constant
  - Assures the posterior integrates to 1

Note on notation

- In class, we will use $p(\cdot)$ for *any* pdf
  - $p(\mathbf{y} \mid \theta)$, $p(\theta)$, $p(\theta, \mathbf{y})$, $p(\theta \mid \mathbf{y})$, $p(\mathbf{y})$, etc.

- For discrete variables, use $P(\cdot)$ for *any* pmf.
- Common alternative notations:
  - π for prior: $\pi(\theta)$
  - f for sampling model / likelihood: $f(\mathbf{y} \mid \theta)$
  - m for marginal distribution: $m(\mathbf{y})$
  - p for anything else. E.g., $p(\theta \mid \mathbf{y})$, $p(\theta, \mathbf{y})$.

Prior and posterior

- If $\mathbf{y}$ is discrete, may write

$$p(\theta \mid \mathbf{y} = \mathbf{k}) = \frac{P(\mathbf{y} = \mathbf{k} \mid \theta) p(\theta)}{\int P(\mathbf{y} = \mathbf{k} \mid \theta) p(\theta) \, d\theta}.$$

- If θ is discrete with possible values Θ, may write

$$P(\theta = k \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta = k) P(\theta = k)}{\sum_{k \in \Theta} p(\mathbf{y} \mid \theta = k) P(\theta = k)}.$$

Example: Sex proneness

- Research suggests the chance of a male or female birth depends on the family.
- The Gupta's (of CNN fame) have three daughters.
- What is the probability their next child will be a daughter?
- A (poor) maximum likelihood solution:
  - Let θ be probability of a daughter for Gupta's
  - Binomial likelihood for first 3 births, with $y{=}\#$ daughters, is

$$P(Y = 3 \mid \theta) = \theta^3$$

  - Maximized at $\hat{\theta} = 1$

Example: Sex proneness

- Alternatively, use prior for θ
- Naive approach: $\theta \sim \text{Uniform}(0,1)$. $= Beta(1,1)$
  - So $p(\theta) = 1$ for $0 \leq \theta \leq 1$.
- What is $p(\theta|y=3)$?

$P(Y=3) = \int P(y=3|\theta) P(\theta) d\theta$

$= \int_0^1 \theta^3 \cdot 1 \, d\theta$

$= \frac{1}{4} - 0 = \frac{1}{4}$

$P(\theta|y=3) = 4 \cdot P(y=3|\theta) P(\theta)$

$= 4 \cdot \theta^3$

- Estimate θ using this posterior.

$E(\theta|y=3) = \int_0^1 \theta \cdot P(\theta|y=3) d\theta = \int_0^1 4 \cdot \theta^4 d\theta$

$= \frac{4}{5} = 0.8$

- Assume $\theta \sim \text{Beta}(a, b)$:

$$p(\theta) = \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

where $B(\cdot, \cdot)$ is the *beta function*.

- If $Y \sim \text{Binomial}(n, \theta)$, then

$$p(\theta | y = k) = \text{Beta}(a + k, b + n - k)$$

$$P(Y=k|\theta) = \binom{n}{k} \theta^{k} (1-\theta)^{n-k}$$

$$P(Y=k|\theta) \cdot P(\theta) = \underbrace{C}_{\text{constant}} \cdot \theta^{a+k-1} (1-\theta)^{b+n-k-1}$$

So, $P(\theta|Y=k) \propto \theta^{a+k-1} (1-\theta)^{b+n-k-1} \propto \text{Beta}(a+k, b+n-k)$

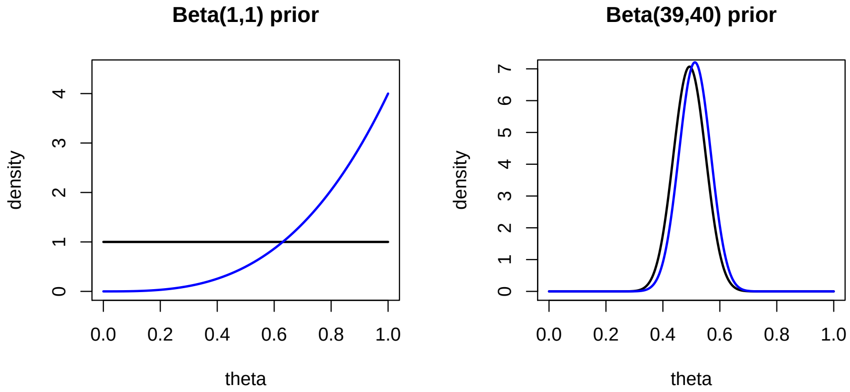
Conclude $P(\theta|Y=k) =$ → $\text{Beta}(a+k, b+n-k)$

Example: Sex proneness (cont.)

- Parent proneness is thought to have a very small effect, if any.
- Based on data from many families, a more realistic prior for θ is
$$p(\theta) = \text{Beta}(39, 40).$$
- For the Gupta family, $p(\theta \mid y = 3) = \text{Beta}(42, 40)$
- The expected value of $\text{Beta}(a, b)$ is $a/(a + b)$:
  - Prior estimate is $E_\theta = 39/79 = 0.494$
  - Posterior estimate is $E_{\theta \mid y=3} = 42/82 = 0.512$

Example: Sex proneness (cont.)

- **Prior** and posterior densities, given $y = 3$:



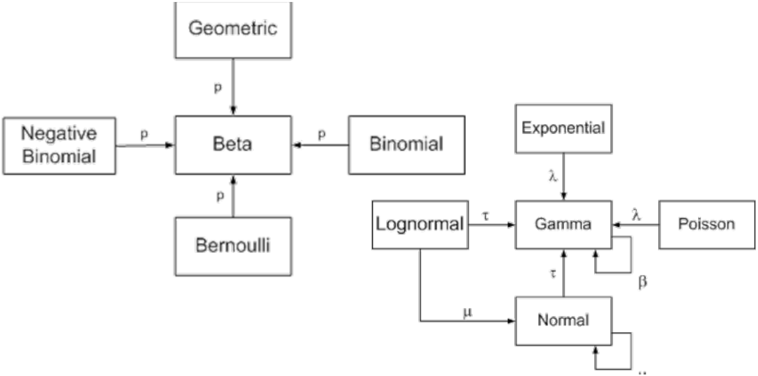
Code: http://www.ericfrazerlock.com/Prior_and_posterior_Rcode1.r

Conjugate priors

- A prior is *conjugate* for a given likelihood if its posterior belongs to the same distributional family.
- A beta prior is conjugate for binomial data
  - Both $p(\theta)$ and $p(\theta \mid y)$ give beta distributions.
- Conjugate priors facilitate computation of posteriors
- Particularly useful when updating the posterior adaptively
  - E.g., after one girl, posterior is $\text{Beta}(40, 40)$
  - After another girl, posterior is Beta $(41, 40)$, etc.
- Otherwise, no profound theoretical justification.

Conjugate priors

- Common conjugate families (credit: John D. Cook)

- 

See http://en.wikipedia.org/wiki/Conjugate_prior

Normal-normal model

$$p(y \mid \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}}$$

- Assume $y \sim \text{Normal}(\mu, \sigma^2)$ with σ^2 known
- If $p(\mu) = \text{Normal}(\mu_0, \tau^2)$, then $p(\mu) = \frac{1}{\tau\sqrt{2\pi}} e^{\left\{-\frac{(\mu-\mu_0)^2}{2\tau^2}\right\}}$

$$p(\mu \mid y) = \text{Normal}\left(\frac{\sigma^2\mu_0 + \tau^2 y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$$

$\leftarrow A \qquad \leftarrow B$

$$p(\mu \mid y) \propto p(y \mid \mu)\, p(\mu)$$

$$\propto \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\tau^2}\right\}$$

$$\ldots = \exp\left\{-\frac{(\mu-A)^2}{2B}\right\}$$

$$\propto N(A, B)$$

Normal-normal model

- Assume $\mathbf{y} = (y_1, \ldots, y_n)$ are iid with $y_i \sim \text{Normal}(\mu, \sigma^2)$, and σ^2 known.
- If $p(\mu) = \text{Normal}(\mu_0, \tau^2)$, then

$$p(\mu \mid \mathbf{y}) = \text{Normal}\left(\frac{\sigma^2\mu_0 + n\tau^2\bar{y}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

where $\bar{y} = \frac{\sum y_i}{n}$.

- Homework.

Example: Coke bottles

- Coca-Cola bottling machines fill with known variance 0.05 oz
- Each machine is calibrated to fill mean capacity μ
- Bottles are filled with Gaussian error: Normal$(\mu, 0.05)$

  $p(y_i \mid \mu)$

- Historical data show machine calibrations are approximately Normal$(12, 0.01)$.

  $p(\mu)$

Example: Coke bottles

- Five randomly selected bottles from a given machine have sample mean $\bar{y} = 11.88$
- What is the posterior for the calibration of this machine?

$\sigma^2 = 0.05$

$\tau^2 = 0.01$

$\mu_0 = 12$

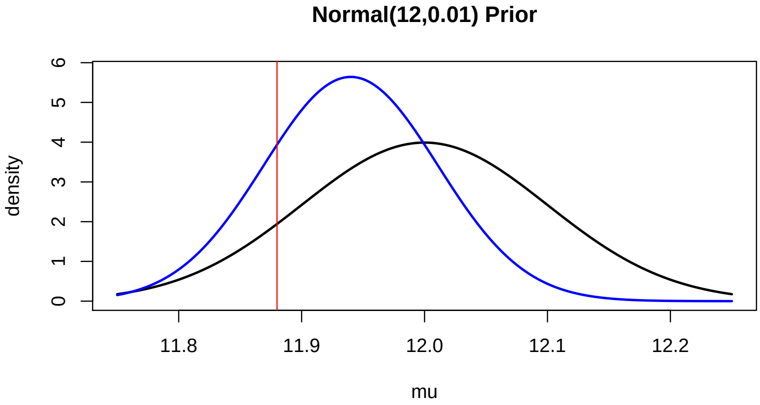
$P(\mu \mid \bar{y}) = N(A, B)$

$A = 11.94 \quad B = 0.005$

- $p(\mu \mid \mathbf{y}) = \text{Normal}(11.94, 0.005)$

Example: Coke bottles

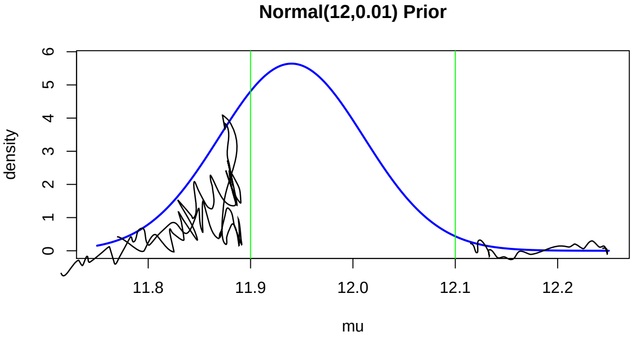
- **Prior** and posterior densities



Code: http://www.ericfrazerlock.com/Prior_and_posterior_Rcode2.r

Example: Coke bottles

- Re-calibrate machines if do not fill within 11.9 and 12.1 oz on average
- What is the probability this machine needs recalibration?



$$\Phi(\frac{11.9 - 11.94}{\sqrt{0.005}}) + 1 - \Phi(\frac{12.1 - 11.94}{\sqrt{0.005}}) \approx 0.30$$

Prior and posterior predictive

- Often we are not interested in making inference on θ, but rather on our best guess for the distribution of y_i
- Can estimate density of the full model by integrating over θ
- The *prior predictive* is the marginal distribution of an observation given the prior:

$$p(y_1) = \int p(y_1 \mid \theta) p(\theta) d\theta$$

- The *posterior predictive* is the distribution for a future observation y_{n+1} given the data so far. If $y_1, \ldots, y_n, y_{n+1} \overset{iid}{\sim} p(y \mid \theta)$,

$$p(y_{n+1} \mid \mathbf{y}) = \int p(y_{n+1} \mid \theta) p(\theta \mid \mathbf{y}) d\theta.$$

$$= \int p(y_{n+1}, \theta \mid \tilde{y}) d\theta$$

Normal-normal predictives

- For the normal-normal model introduced earlier, the prior predictive is

$$p(y_i) = \text{Normal}\left(\mu_0, \tau^2 + \sigma^2\right)$$

- The posterior predictive is

$$p(y_{n+1} \mid \mathbf{y}) = \text{Normal}\left(\frac{\sigma^2\mu_0 + n\tau^2\bar{y}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2} + \sigma^2\right)$$

- Homework

Example: Coke bottles (cont)

- The **predictive for a single bottle from a given machine** is

$$\text{Normal}(12, 0.06)$$

- The predictive for the sixth bottle after the five observed is

$$\text{Normal}(11.94, 0.055)$$