Lecture 4

Multi-Parameter Models (II); Simple Simulation Scheme; Asymptotics

Textbook Ch3.6; Ch3.7; Ch4

邓婉璐

wanludeng@tsinghua.edu.cn

Outline

- Multi-parameter models (II): Multivariate normal
- Simple Simulation Scheme
- Asymptotics
  - Approximation
  - Interpretation (vs Frequentist)
  - Non-applicable scenarios

Objectives for Today

- 类比一元，理解和掌握多元正态分布的先验确定与后验推断

- 理解可以通过抽样获取后验分布的相关信息
- 掌握简单抽样方法(离散格点法)；利用变换为一元的技巧

- 理解正态分布近似的意义，了解其用途
- 掌握如何求得正态分布近似
- 通过理解假设，了解正态分布近似的适用场景

Multivariate Normal Model

- Conjugate prior
- Noninformative prior

Univariate Normal with a Conjugate Prior

Review

Univariate normal with unknown mean & variance:

$$y|\mu,\sigma^2 \sim \mathrm{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Conjugate prior:

$$p(\mu,\sigma^2) \propto p(\sigma^2)p(\mu|\sigma^2) \dashrightarrow \text{N-Inv-}\chi^2(\mu_0,\sigma_0^2/\kappa_0;\nu_0,\sigma_0^2)$$

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0,\sigma_0^2) \qquad \mu|\sigma^2 \sim \mathrm{N}(\mu_0,\sigma^2/\kappa_0)$$

$$p(\mu,\sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2]\right)$$

Joint posterior:

$$p(\mu,\sigma^2|y) = \text{N-Inv-}\chi^2(\mu_n,\sigma_n^2/\kappa_n;\nu_n,\sigma_n^2),$$

Conditional posterior:

$$\mu|\sigma^2,y \sim \mathrm{N}(\mu_n,\sigma^2/\kappa_n) = \mathrm{N}\left(\frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}}\right)$$

Marginal posterior:

$$\sigma^2|y \sim \text{Inv-}\chi^2(\nu_n,\sigma_n^2)$$

$$p(\mu|y) \propto \left(1+\frac{\kappa_n(\mu-\mu_n)^2}{\nu_n\sigma_n^2}\right)^{-(\nu_n+1)/2} = t_{\nu_n}(\mu|\mu_n,\sigma_n^2/\kappa_n).$$

Wishart and Inverse Wishart Distributions

Wishart distribution ($Wishart_\nu(\Lambda)$)

- Multivariate analogue of a scaled χ^2 distribution
- If $z_1, \ldots, z_\nu \sim^{iid} N_d(0, \Lambda)$ then

$$\Sigma = \sum_{i=1}^{\nu} z_i z_i^T \sim Wishart_\nu(\Lambda)$$

- Like $z_1, \ldots, z_\nu \sim^{iid} N(0, \tau^2)$ then

$$S = \sum_{i=1}^{\nu} z_i^2 \sim \tau^2 \chi_\nu^2$$

Inverse Wishart distribution (Inv-Wishart$_\nu(\Lambda^{-1})$)

- Multivariate analogue of a scaled Inv $-\chi^2$ distribution
- If $\Sigma \sim \text{Wishart}_\nu(\Lambda)$ then

$\Sigma^{-1} \sim \text{Inv} - \text{Wishart}_\nu(\Lambda)$ 教材的记号

$\Sigma^{-1} \sim \text{Inv} - \text{Wishart}_\nu(\Lambda^{-1})$ R, Python等的记号

注意：常见有上下两种表达形式，其PDF是一致的。

Multivariate Normal with Unknown Mean & Variance

Likelihood: $p(y_1,\ldots,y_n|\mu,\Sigma) \propto |\Sigma|^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right)$ $y \in \mathbb{R}^d$

$= |\Sigma|^{-n/2}\exp\left(-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S_0)\right)$, $S_0 = \sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)^T$

Conjugate prior: $p(\mu,\Sigma) = p(\Sigma)p(\mu \mid \Sigma)$ Normal-Inverse-Wishart$(\mu_0, \Lambda_0/\kappa_0; \nu_0, \Lambda_0)$

$\Sigma \sim \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1})$ $\mu|\Sigma \sim \mathrm{N}(\mu_0, \Sigma/\kappa_0)$

$p(\mu,\Sigma) \propto |\Sigma|^{-((\nu_0+d)/2+1)}\exp\left(-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right)$

Joint posterior: $(\mu,\Sigma) \mid y \sim$ Normal-Inverse-Wishart $(\mu_n, \Lambda_n/\kappa_n; \nu_n, \Lambda_n)$

$\mu|\Sigma, y \sim \mathrm{N}(\mu_n, \Sigma/\kappa_n)$

$\Sigma|y \sim \text{Inv-Wishart}_{\nu_n}(\Lambda_n^{-1})$

$\mu|y \sim t_{\nu_n-d+1}(\mu_n, \Lambda_n/(\kappa_n(\nu_n-d+1)))$

Calculation for Parameters of Posterior

$$
\begin{aligned}
p(\mu, \Sigma \mid y) &\propto p(y \mid \mu, \Sigma) \cdot p(\mu, \Sigma) \\
&\propto |\Sigma|^{-((n+\nu_0+d)/2+1)} \exp\Big(-\frac{1}{2} tr(\Sigma^{-1} S_0) - \frac{\kappa_0}{2}(\mu-\mu_0)^T \Sigma^{-1}(\mu-\mu_0) - \frac{1}{2} tr(\Lambda_0 \Sigma^{-1}) \Big)
\end{aligned}
$$

$$\nu_n = \nu_0 + n$$

$$\kappa_0(\mu_0^T\Sigma^{-1}\mu_0 - 2\mu_0^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu)$$

$$
\begin{aligned}
tr(\Sigma^{-1}S_0) &= \sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu) \\
&= \sum_{i=1}^{n}(y_i-\bar{y}+\bar{y}-\mu)^T\Sigma^{-1}(y_i-\bar{y}+\bar{y}-\mu) \\
&= \sum_{i=1}^{n}(y_i-\bar{y})^T\Sigma^{-1}(y_i-\bar{y}) - 2\sum_{i=1}^{n}(y_i-\bar{y})^T\Sigma^{-1}(\bar{y}-\mu) + n(\bar{y}-\mu)^T\Sigma^{-1}(\bar{y}-\mu) \\
&= \sum_{i=1}^{n}(y_i-\bar{y})^T\Sigma^{-1}(y_i-\bar{y}) + n(\mu-\bar{y})^T\Sigma^{-1}(\mu-\bar{y}) \\
&= tr(\Sigma^{-1}S) + n(\mu-\bar{y})^T\Sigma^{-1}(\mu-\bar{y})
\end{aligned}
$$

$$n(\bar{y}^T\Sigma^{-1}\bar{y} - 2\bar{y}\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu)$$

$$S = \sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})^T$$

Calculation for Parameters of Posterior

Prior: $p(\mu,\Sigma)\propto|\Sigma|^{-((\nu_0+d)/2+1)}\exp\left(-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1})-\frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right)$

$$
\begin{aligned}
& tr(\Sigma^{-1}S_0)+\kappa_0(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)+tr(\Lambda_0\Sigma^{-1}) \\
= \;& (n+\kappa_0)\mu^T\Sigma^{-1}\mu-2(n\bar{y}+\kappa_0\mu_0)^T\Sigma^{-1}\mu+(n\bar{y}^T\Sigma^{-1}\bar{y}+\kappa_0\mu_0^T\Sigma^{-1}\mu_0)+\; tr(\Sigma^{-1}S)+tr(\Lambda_0\Sigma^{-1}) \\
= \;& (n+\kappa_0)\left(\mu-\frac{n\bar{y}+\kappa_0\mu_0}{n+\kappa_0}\right)^T\Sigma^{-1}\left(\mu-\frac{n\bar{y}+\kappa_0\mu_0}{n+\kappa_0}\right)-(n+\kappa_0)\left(\frac{n\bar{y}+\kappa_0\mu_0}{n+\kappa_0}\right)^T\Sigma^{-1}\left(\frac{n\bar{y}+\kappa_0\mu_0}{n+\kappa_0}\right) \\
& +(n\bar{y}^T\Sigma^{-1}\bar{y}+\kappa_0\mu_0^T\Sigma^{-1}\mu_0)+\; tr(\Sigma^{-1}S)+tr(\Lambda_0\Sigma^{-1}) \\
= \;& (n+\kappa_0)\left(\mu-\frac{n\bar{y}+\kappa_0\mu_0}{n+\kappa_0}\right)^T\Sigma^{-1}\left(\mu-\frac{n\bar{y}+\kappa_0\mu_0}{n+\kappa_0}\right) \\
& +tr\left(\left[-\frac{1}{n+\kappa_0}(n\bar{y}+\kappa_0\mu_0)(n\bar{y}+\kappa_0\mu_0)^T+n\bar{y}\bar{y}^T+\kappa_0\mu_0\mu_0^T+S+\Lambda_0\right]\Sigma^{-1}\right)
\end{aligned}
$$

$\Rightarrow \quad \mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0+\frac{n}{\kappa_0+n}\bar{y}$

$\kappa_n = \kappa_0+n$

$$
\begin{aligned}
\Lambda_n &= -\frac{1}{n+\kappa_0}(n\bar{y}+\kappa_0\mu_0)(n\bar{y}+\kappa_0\mu_0)^T+n\bar{y}\bar{y}^T+\kappa_0\mu_0\mu_0^T+S+\Lambda_0 \\
&= \frac{n\kappa_0}{n+\kappa_0}(\bar{y}-\mu_0)(\bar{y}-\mu_0)^T+S+\Lambda_0
\end{aligned}
$$

Comparison between 1-dim and d-dim

$$p(\mu,\sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2]\right)$$

Meaning of the 4 hyper-parameters:

μ_0	Prior mean
σ_0^2	Prior sample variance
κ_0	# of additional data for prior mean
ν_0	# of additional data for prior variance

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2.
\end{aligned}$$

$$p(\mu,\Sigma) \propto |\Sigma|^{-((\nu_0+d)/2+1)} \exp\left(-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right)$$

Meaning of the 4 hyper-parameters:

μ_0	Prior mean
Λ_0	Prior sum of squares
κ_0	# of additional data for prior mean
ν_0	# of additional data for prior covariance matrix

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\Lambda_n &= \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)(\bar{y}-\mu_0)^T
\end{aligned}$$

$$S = \sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})^T$$

Multivariate Normal Model

- Conjugate prior
- Noninformative prior

Univariate Normal with a Conjugate Prior

Review

Univariate normal with unknown mean & variance:

$$y|\mu,\sigma^2 \sim \mathrm{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Conjugate prior:

$$p(\mu,\sigma^2) \propto p(\sigma^2)p(\mu|\sigma^2) \dashrightarrow \text{N-Inv-}\chi^2(\mu_0,\sigma_0^2/\kappa_0;\nu_0,\sigma_0^2)$$

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0,\sigma_0^2) \qquad \mu|\sigma^2 \sim \mathrm{N}(\mu_0,\sigma^2/\kappa_0)$$

$$p(\mu,\sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2]\right)$$

Non-informative prior:

$$\kappa_0 \to 0, \nu_0 \to -1, \sigma_0^2 \to 0$$

$$p(\mu,\sigma^2) \propto (\sigma^2)^{-1} \dashrightarrow$$

uniform on $(\mu, \log\sigma)$ ⟸ Jeffreys' s principle
prior independence of location and scale parameters

Multivariate Normal with Unknown Mean & Variance

Similar to the Univariate normal case:

$$p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$$

multivariate Jeffreys' prior

Assume prior independence of location and scale parameters

the limit of the conjugate prior as $\kappa_0 \to 0$, $\nu_0 \to -1$, and $|\Lambda_0| \to 0$.

The posterior in this case satisfies

$$\mu|\Sigma, y \sim N(\bar{y}, \frac{\Sigma}{n})$$

$$\Sigma|y \sim Inv - Wishart_{n-1}(S^{-1})$$

$$\mu|y \sim t_{n-d}(\bar{y}, \frac{S}{n(n-d)})$$

Simple Simulation Scheme

Example: Air Conditioning Failures in a Boeing 720

- There are 13 planes in the complete dataset. (Proschan, 1963)
- For the plane 7910, the times between failures (in hours) are 74, 57, 48, 29, 502, 12, 70, 21, 29, 386, 59, 27, 153, 26, 326 (n = 15).

- What is the average time between failures?

- Ref: Proschan, F. (1963). Theoretical Explanation of Observed Decreasing Failure Rate. *Technometrics*, *5*(3), 375–383. https://doi.org/10.2307/1266340

Example: Air Conditioning Failures in a Boeing 720

- Y: the times between failures (in hours) are 74, 57, 48, 29, 502, 12, 70, 21, 29, 386, 59, 27, 153, 26, 326 (n = 15).
- Assume that $y_i|\theta \sim Exp(\theta)$, where $\theta = \frac{1}{E[y|\theta]}$ is often referred to as the rate parameter (i.e. number of events per unit time).

$$p(y|\theta) = \prod_{i=1}^{n} \theta e^{-y_i\theta} = \theta^n e^{-n\bar{y}\theta}$$

which gives the MLE as $\hat{\theta} = \frac{1}{\bar{y}}$

- $\hat{\theta} = 0.00825$ (about 8 failures for every 1000 hours of flight time). Thus the average time between failures is $\bar{y} = 121.27$ hours.

Example: Air Conditioning Failures in a Boeing 720

Bayesian Solution:

$$p(y|\theta) = \prod_{i=1}^{n} \theta e^{-y_i\theta} = \theta^n e^{-n\bar{y}\theta}$$

- A conjugate prior for the exponential distribution is the gamma.
- If $\theta \sim \Gamma(\alpha, \beta)$, then the posterior is

$$p(\theta|y) \propto \theta^n e^{-n\bar{y}\theta} \theta^{\alpha-1} e^{-\beta\theta} = \theta^{\alpha+n-1} e^{-(\beta+n\bar{y})\theta}$$

  i.e. $\Gamma(\alpha + n, \beta + n\bar{y})$.

- This gamma prior can be thought of as $\alpha - 1$ exponential observations totalling β.

Example: Air Conditioning Failures in a Boeing 720

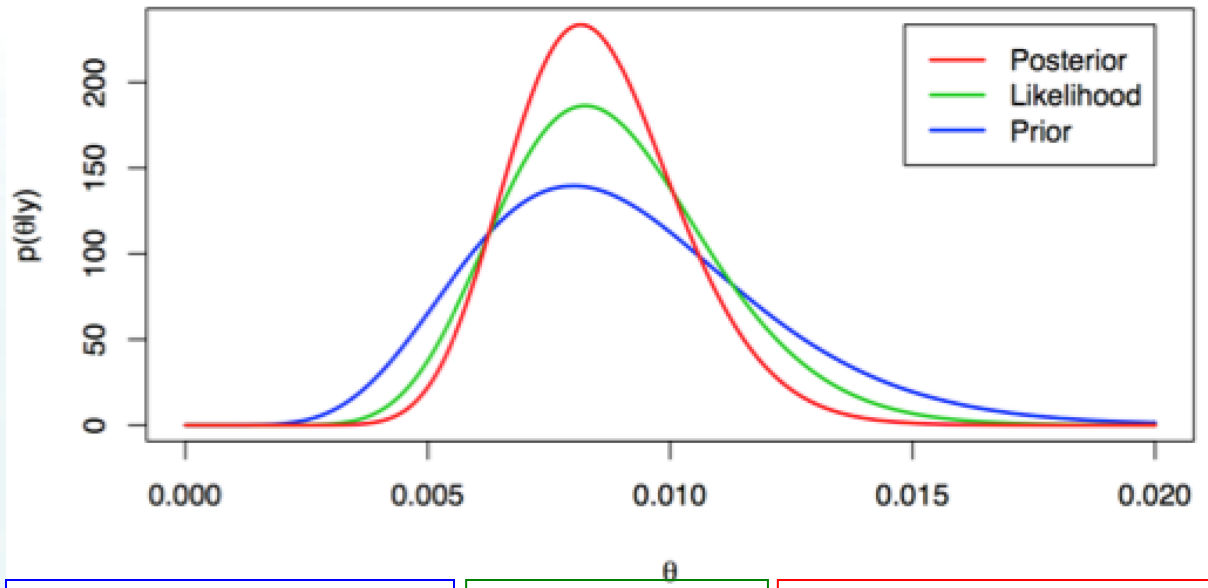
$\theta \sim \Gamma(\alpha, \beta)$. What hyperparameter for this example? i.e. $\alpha, \beta =?$

- For the example, we can use some of the other planes to develop a gamma prior for plane 7910.
- Here we took 4 of the planes and calculated the MLEs of θ for each of them. The average of these was about 0.009 with a standard deviation of 0.003.
- $\Gamma(\alpha, \beta)$ with this mean and standard deviation has $\alpha = 9, \beta = 1000$.
- Let's also use a less informative prior to see how dependent on our answer is on our prior choice.
- Since the above prior corresponds to 8 observations, lets use corresponding to half as many observations ($\alpha = 5$), with half as much time ($\beta = 500$)

Example: Air Conditioning Failures in a Boeing 720

$$n = 15, \bar{y} = 121.27, \alpha = 9, \beta = 1000$$

Strong prior

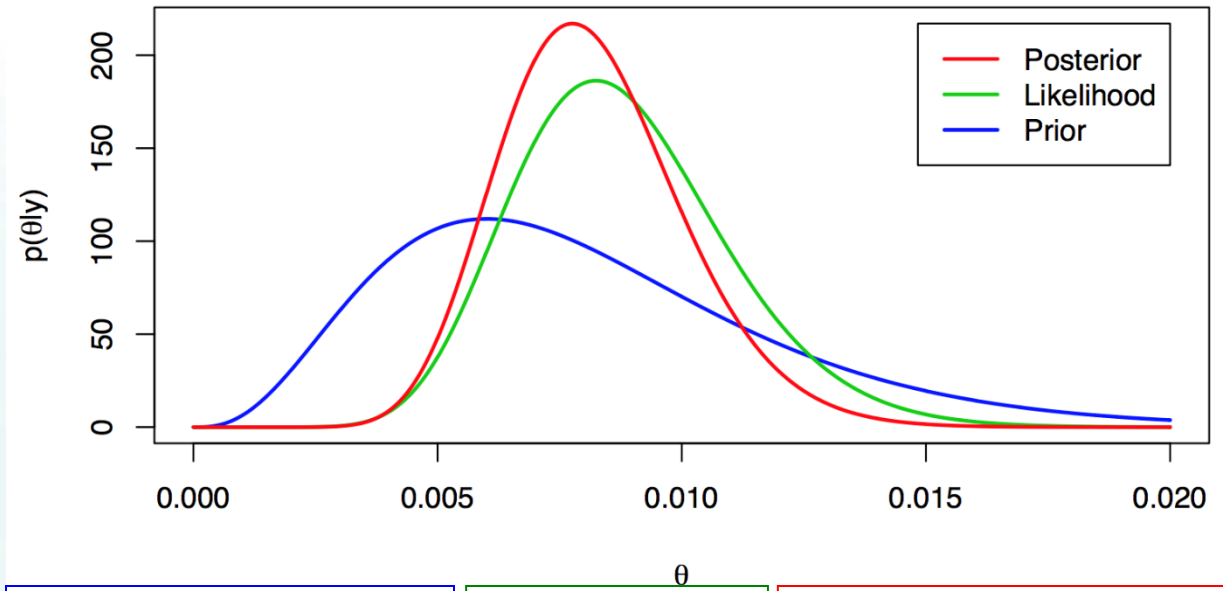


Prior	Likelihood	Posterior
$E(\theta) = 0.009$ $\mathrm{Var}(\theta) = 0.000009$ $\mathrm{SD}(\theta) = 0.003$	$\hat{\theta} = 0.00825$	$E(\theta\|y) = 0.00851$ $\mathrm{Var}(\theta\|y) = 0.000003$ $\mathrm{SD}(\theta\|y) = 0.00173$

Example: Air Conditioning Failures in a Boeing 720

$$n = 15, \bar{y} = 121.27, \alpha = 5, \beta = 500$$

Weak prior



Prior	Likelihood	Posterior
$E(\theta) = 0.01$ $\mathrm{Var}(\theta) = 0.00002$ $\mathrm{SD}(\theta) = 0.0045$	$\hat{\theta} = 0.00825$	$E(\theta\|y) = 0.00862$ $\mathrm{Var}(\theta\|y) = 0.0000037$ $\mathrm{SD}(\theta\|y) = 0.00193$

Function of Parameters

- What is the standard deviation of time between failures?

- Actually we can answer much more (complicated) questions by simulation.

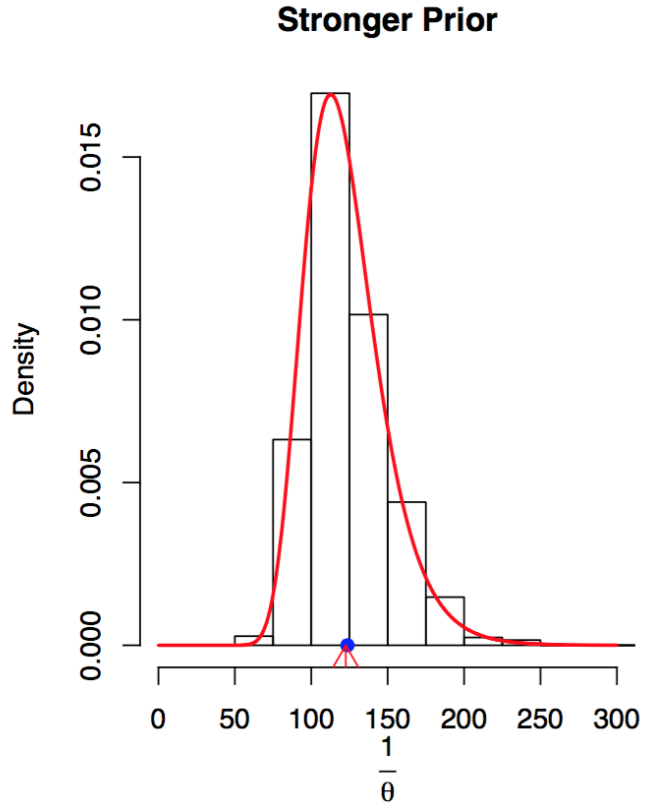
- Let $\theta_i \sim p(\theta|y)$ $i.i.d.$, $i = 1, \ldots, m$, and suppose we are interested in $\lambda = h(\theta)$ for some function $h(\cdot)$. Then

$$h(\theta_i) = \lambda_i \sim p(\lambda|y) \; i.i.d., \; i = 1, \ldots, m$$

- E.g., take $h(\theta) = \frac{1}{\theta}$. Then $\bar{\lambda}$ is an unbiased estimate of $E[\frac{1}{\theta} | \, y]$.

Example: Air Conditioning Failures in a Boeing 720

- Suppose that we are interested in $\lambda = \frac{1}{\theta} = E[y|\theta]$, the expected time between failures.
- In fact, we know that $\lambda|y$ has an inverse gamma distribution in this example.



Prior	$E\left[\frac{1}{\theta}\|y\right]$	$\hat{E}\left[\frac{1}{\theta}\|y\right]$	$\mathrm{SD}\left(\frac{1}{\theta}\|y\right)$	$\widehat{\mathrm{SD}}\left(\frac{1}{\theta}\|y\right)$
$\alpha = 9, \beta = 1000$	122.56	123.68	26.13	27.00

Univariate Normal with a Conjugate Prior

Review

Univariate normal with unknown mean & variance:

$$y|\mu,\sigma^2 \sim \mathrm{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Conjugate prior:

$$p(\mu,\sigma^2) \propto p(\sigma^2)p(\mu|\sigma^2) \dashrightarrow \text{N-Inv-}\chi^2(\mu_0,\sigma_0^2/\kappa_0;\nu_0,\sigma_0^2)$$

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0,\sigma_0^2) \qquad \mu|\sigma^2 \sim \mathrm{N}(\mu_0,\sigma^2/\kappa_0)$$

Joint posterior:

$$p(\mu,\sigma^2|y) = \text{N-Inv-}\chi^2(\mu_n,\sigma_n^2/\kappa_n;\nu_n,\sigma_n^2),$$

Generating from a Joint Distribution

- Want to simulate X, Y from $p(x, y)$:
  - Sample x_i from $p(x), i = 1, \ldots, m$.
  - Sample y_i from $p(y|x_i), i = 1, \ldots, m$.

- Justification that this scheme actually draws from the joint distribution:
- The joint empirical CDF (ECDF) of $\{(x_i, y_i)\ i = 1, \ldots, m\}$ is

$$\hat{P}(x, y) = \frac{1}{m} \sum_{i=1}^{m} I(x_i \leq x, y_i \leq y) = \frac{1}{m} \sum_{i=1}^{m} I(x_i \leq x) I(y_i \leq y)$$

Generating from a Joint Distribution

- The expected value of the ECDF is

$$E[\hat{P}(x,y)] = P[X \le x, Y \le y] = P(x,y)$$

since $E[I(x_i \le x)I(y_i \le y)]$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x_i \le x)I(y_i \le y)p(x_i)p(y_i|x_i)dy_i dx_i$$

$$= \int_{-\infty}^{x}\int_{-\infty}^{y} p(x_i, y_i)dy_i dx_i$$

$$= P[X \le x, Y \le y]$$

- The ECDF is an unbiased estimate of the CDF, and converge to the CDF by LLN.

$$\hat{P}(x,y) \to P(x,y) \quad \text{as} \quad n \to \infty$$

$$\mathrm{Var}\left(\hat{P}(x,y)\right) = \frac{P(x,y)(1-P(x,y))}{n} \to 0$$

Univariate Normal with a Conjugate Prior

Review

Univariate normal with unknown mean & variance:

$$y|\mu,\sigma^2 \sim \mathrm{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Conjugate prior:

$$p(\mu,\sigma^2) \propto p(\sigma^2)p(\mu|\sigma^2) \dashrightarrow \text{N-Inv-}\chi^2(\mu_0,\sigma_0^2/\kappa_0;\nu_0,\sigma_0^2)$$

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0,\sigma_0^2) \qquad \mu|\sigma^2 \sim \mathrm{N}(\mu_0,\sigma^2/\kappa_0)$$

Joint posterior:

$$p(\mu,\sigma^2|y) = \text{N-Inv-}\chi^2(\mu_n,\sigma_n^2/\kappa_n;\nu_n,\sigma_n^2),$$

Marginal posterior:

$$\sigma^2|y \sim \text{Inv-}\chi^2(\nu_n,\sigma_n^2)$$

Univariate Normal with a Conjugate Prior

Review

Univariate normal with unknown mean & variance:

$$y|\mu,\sigma^2 \sim \mathrm{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Conjugate prior:

$$p(\mu,\sigma^2) \propto (p(\sigma^2))^{(2)-1} p(\mu|\sigma^2) \dashrightarrow \text{N-Inv-}\chi^2(\mu_0, \sigma_0^2/\kappa_0; \nu_0, \sigma_0^2)$$

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0,\sigma_0^2) \qquad \mu|\sigma^2 \sim \mathrm{N}(\mu_0, \sigma^2/\kappa_0)$$

Joint posterior:

$$p(\mu,\sigma^2|y) = \text{N-Inv-}\chi^2(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2),$$

Conditional posterior:

$$\begin{aligned}\mu|\sigma^2, y &\sim \mathrm{N}(\mu_n, \sigma^2/\kappa_n) \\ &= \mathrm{N}\left(\frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}}\right)\end{aligned}$$

Marginal Distribution

- Want to simulate Y based on $p(x, y)$:
  - Sample x_i from $p(x), i = 1, \ldots, m$.
  - Sample y_i from $p(y|x_i), i = 1, \ldots, m$.
  - Keep only y_i , $i = 1, \ldots, m$.

- Justification that this scheme actually draws from the marginal distribution:
- The empirical CDF of $\{y_i, i = 1, \ldots, m\}$ is

$$\hat{P}(y) = \frac{1}{m} \sum_{i=1}^{m} I(y_i \leq y)$$

Marginal Distribution

- The expected value of the ECDF is

$$E[\hat{P}(y)] = P[Y \le y] = P(y)$$

since

$$E[I(y_i \le y)]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(y_i \le y)p(x_i)p(y_i|x_i)dy_i dx_i$$

$$= \int_{-\infty}^{y}\int_{-\infty}^{\infty} p(x_i, y_i)dx_i dy_i$$

$$= P[Y \le y]$$

- The ECDF is an unbiased estimate of the CDF, and converge to the CDF by LLN.

$$\hat{P}(y) \to P(y)$$

$$\mathrm{Var}\left(\hat{P}(y)\right) = \frac{P(y)(1-P(y))}{n} \to 0 \quad \text{as} \quad n \to \infty$$

Univariate Normal

Univariate normal with unknown mean & variance:

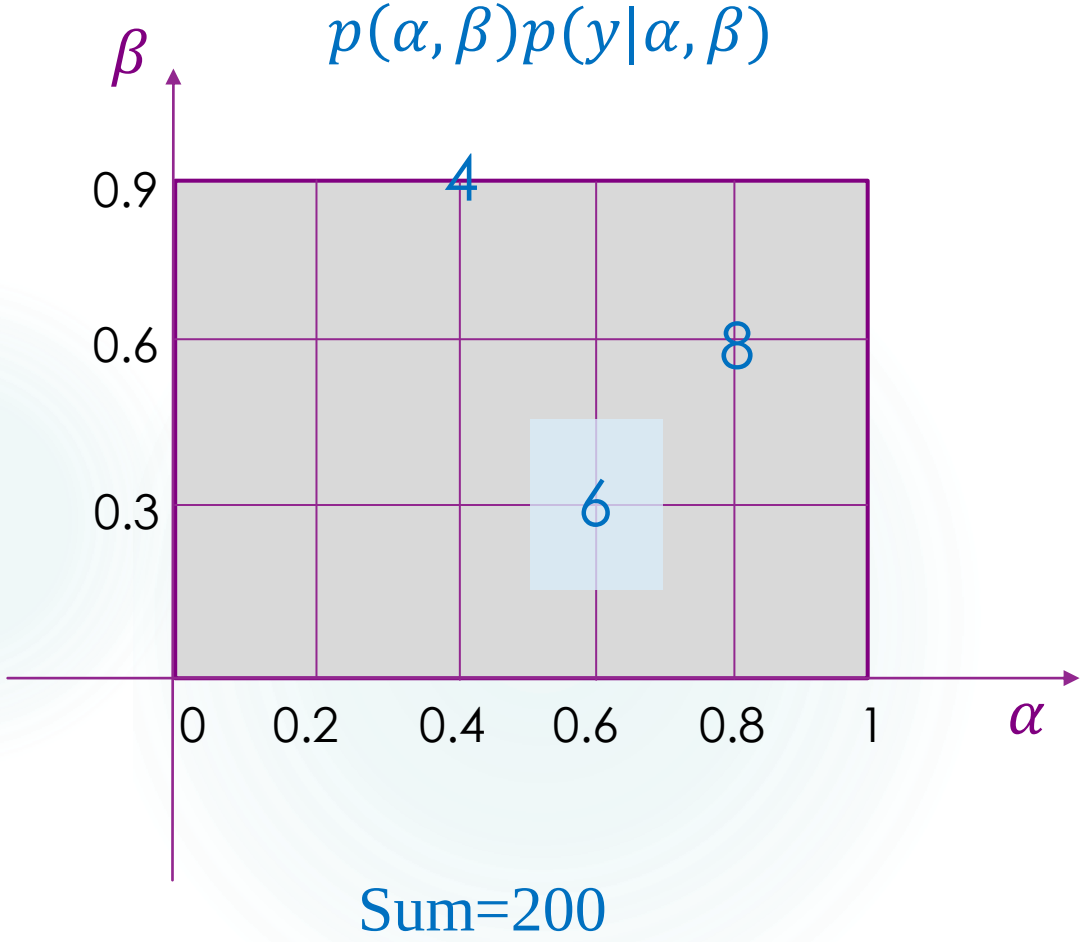
$$y|\mu,\sigma^2 \sim \mathrm{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

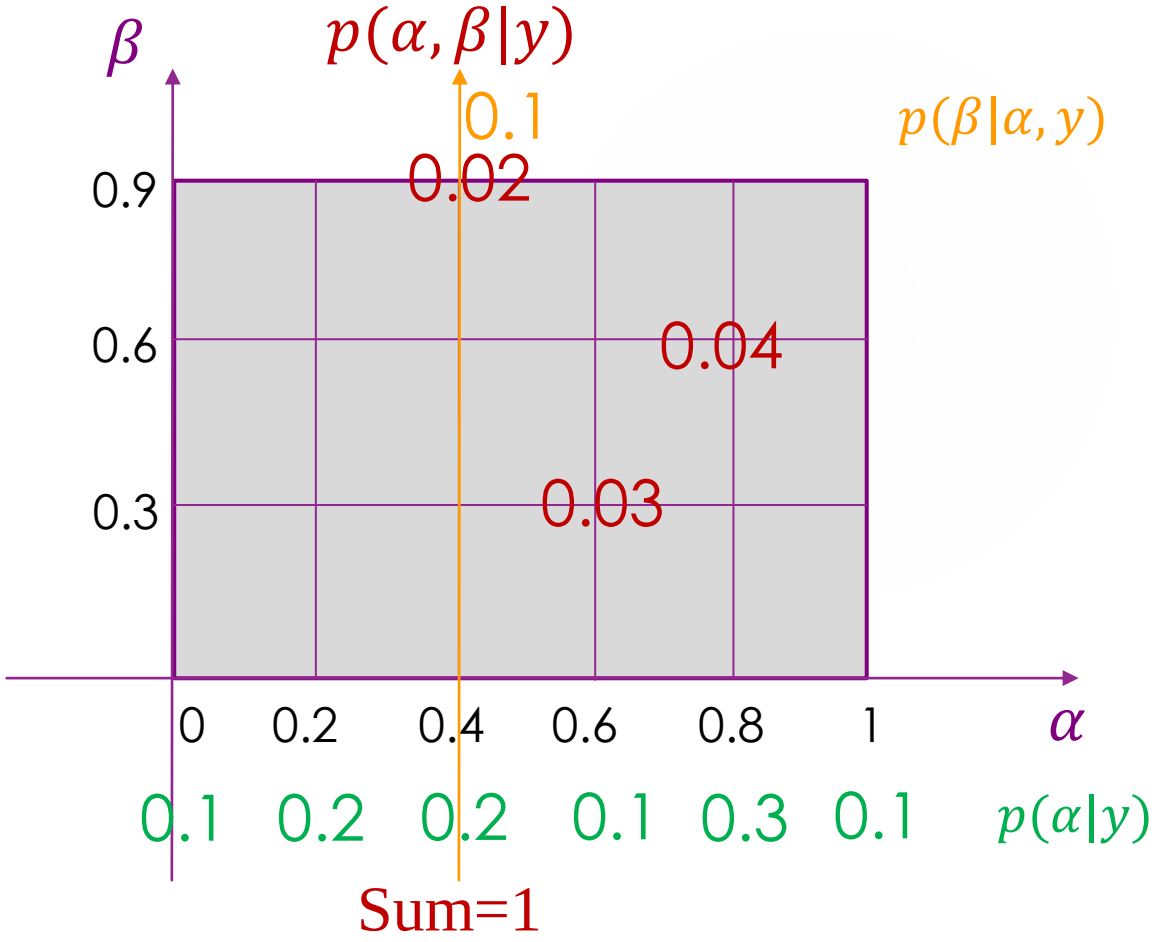
Non-conjugate prior: $p(\mu,\sigma^2) \propto p(\sigma^2)p(\mu|\sigma^2)$

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0,\sigma_0^2) \qquad \mu|\sigma^2 \sim N(\mu_0,\tau_0^2)$$

How to get the samples from posterior?

Example for Discrete-grid sampling procedure





Discrete-grid sampling procedure

- Draw 1000 random samples from the joint posterior distribution of (α, β):

1. After some experimentation, we choose a range that captures almost all the mass of the posterior distribution.
2. Compute the unnormalized posterior density at a grid of values that cover the effective range.
3. Normalize the total probability in the grid to 1.
4. Compute the marginal posterior distribution of α by numerically summing over β in the discrete distribution computed on the grid.
5. For s = 1,...,1000:
   - (a) Draw α^s from the discretely computed $p(\alpha|y)$; this can be viewed as a discrete version of the inverse cdf method.
   - (b) Draw β^s from the discrete conditional distribution, $p(\beta|\alpha, y)$, given the just-sampled value of α.
   - (c) For each of the sampled α and β, add a uniform random jitter centered at zero with a width equal to the spacing of the sampling grid. This gives the simulation draws a continuous distribution.

Asymptotics

- Introduction
- Examples
- Advantages
- Large Sample Theory
- Bayesian vs Frequentist

Asymptotics & Bayesian

- What is asymptotics?
  - Properties of something when sample size n goes to infinite

- Why do we want to discuss asymptotics here?
  - Statistical inference & asymptotics
  - Bayesian inference & asymptotics
  - Bayesian vs frequentist

Normal Approximations to the Posterior Distribution

Unimodal and roughly symmetric

Taylor series expansion of $\log p(\theta|y)$ centered at the posterior mode $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \cdots$$

The linear term: $+\left[\frac{d}{d\theta} \log p(\theta|y)\right]_{\theta=\hat{\theta}}^T (\theta - \hat{\theta})$

zero

Normal approximation: $p(\theta|y) \approx \mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1})$

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$ ----→ observed information

Example: Binomial Distribution

Posterior with the conjugate prior $\theta \sim Beta(\alpha, \beta)$:

$$P(\theta|y) \propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}$$

$$\frac{d \log p(\theta|y)}{d\theta} = \frac{\alpha+y-1}{\theta} - \frac{\beta+n-y-1}{1-\theta}$$

$$\frac{d^2 \log p(\theta|y)}{d\theta^2} = -\frac{\alpha+y-1}{\theta^2} - \frac{\beta+n-y-1}{(1-\theta)^2}$$

$$\hat{\theta} = \frac{\alpha+y-1}{\alpha+\beta+n-2} \qquad I(\hat{\theta}) = \frac{\alpha+\beta+n-2}{\hat{\theta}(1-\hat{\theta})}$$

Normal approximation:

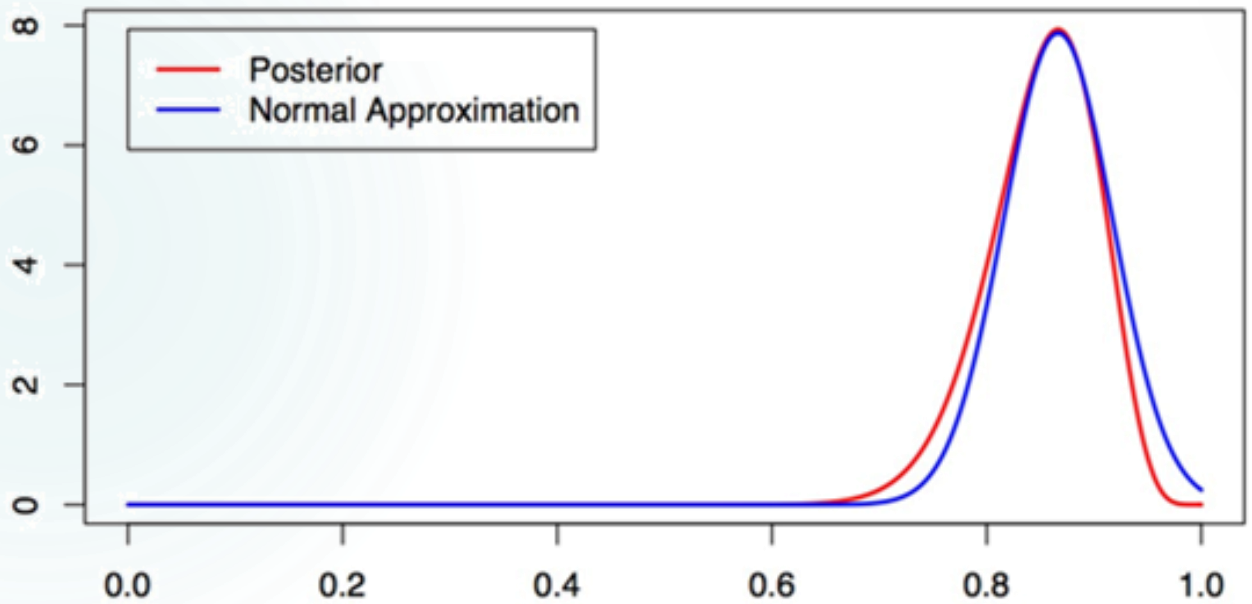
$$p(\theta|y) \approx N\left(\hat{\theta}, \frac{\hat{\theta}(1-\hat{\theta})}{\alpha+\beta+n-2}\right)$$

Example: Binomial Distribution

- Data example: n = 41, y = 37, and a Beta(3, 3) prior

$$\hat{\theta} = \frac{3 + 37 - 1}{6 + 41 - 2} = \frac{39}{45} = 0.8667$$

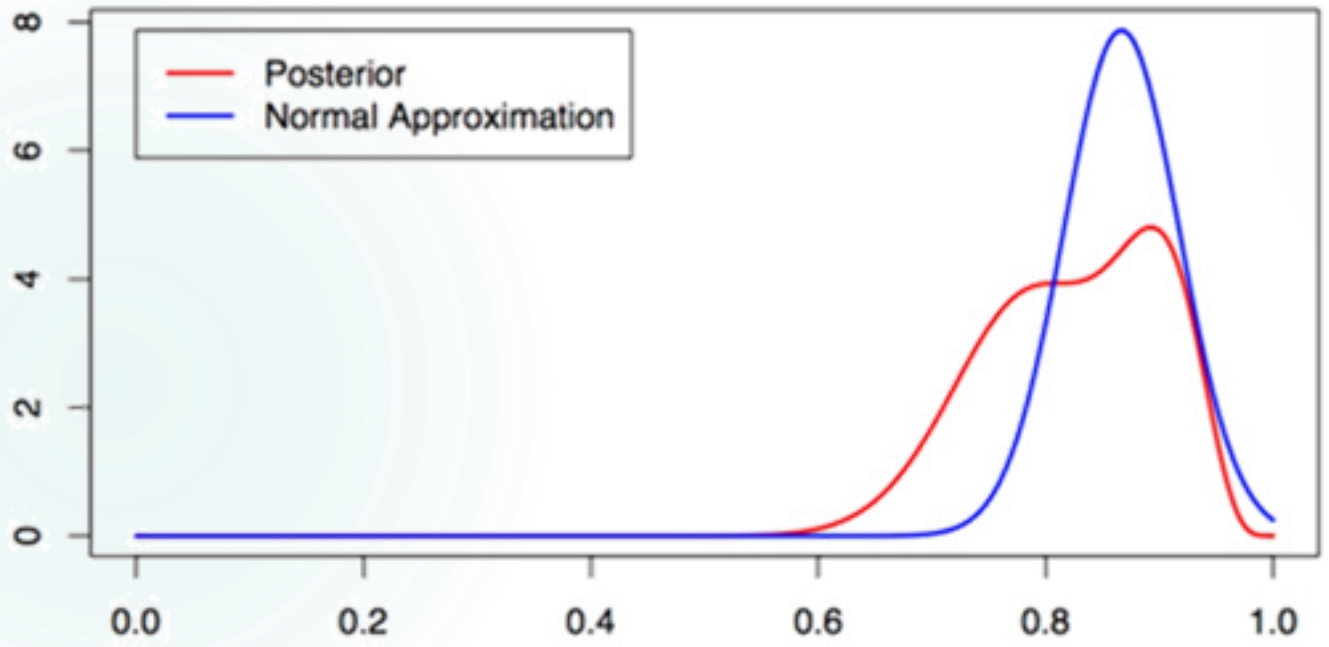
$$I(\hat{\theta}) = \frac{45^3}{39 \times 6} = 389.42 \qquad [I(\hat{\theta})]^{-1/2} = 0.0507$$



Example: Binomial Distribution

Now with the $\frac{1}{2}\text{Beta}(8,2) + \frac{1}{2}\text{Beta}(2,8)$ mixture prior

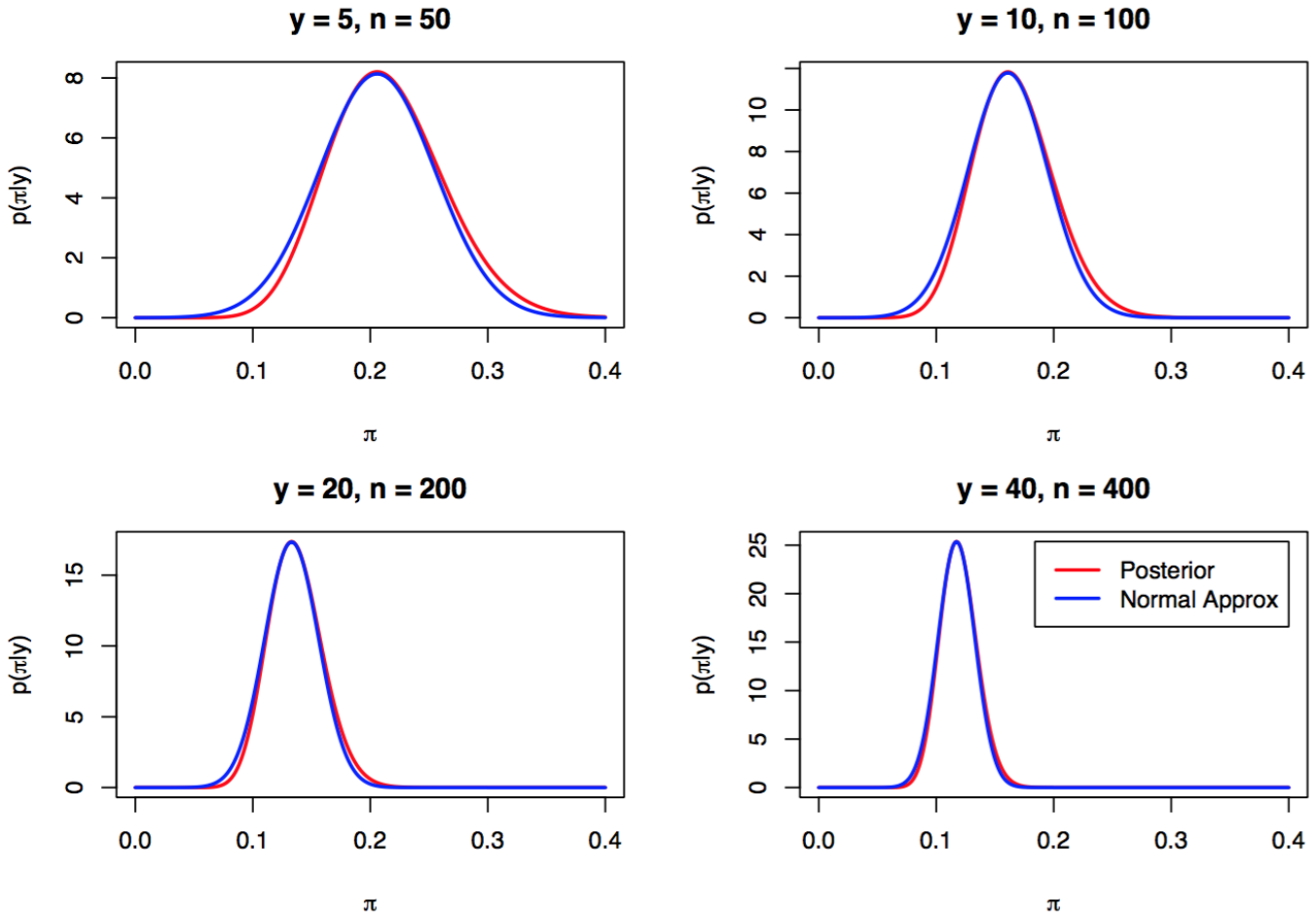
$$\hat{\theta} = 0.8980 \quad I(\hat{\theta}) = 490.11 \quad [I(\hat{\theta})]^{-1/2} = 0.0452$$



The comment about unimodal and symmetric is important.
However when the number of observations gets big, this usually isn't a problem.

Example: Binomial Distribution

Now with the Beta(10,10) prior



Example: Normal Distribution with Unknown Mean & Variance

Posterior with uniform prior density for $(\mu, \log\sigma)$

$$\log p(\mu, \log\sigma|y) = \text{constant} - n\log\sigma - \frac{1}{2\sigma^2}((n-1)s^2 + n(\bar{y}-\mu)^2) \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$\begin{aligned}
\frac{d}{d\mu}\log p(\mu, \log\sigma|y) &= \frac{n(\bar{y}-\mu)}{\sigma^2}, \\
\frac{d}{d(\log\sigma)}\log p(\mu, \log\sigma|y) &= -n + \frac{(n-1)s^2 + n(\bar{y}-\mu)^2}{\sigma^2},
\end{aligned}$$

$$\Rightarrow (\hat{\mu}, \log\hat{\sigma}) = \left(\bar{y}, \log\left(\sqrt{\frac{n-1}{n}}\, s\right)\right)$$

$$\begin{aligned}
\frac{d^2}{d\mu^2}\log p(\mu, \log\sigma|y) &= -\frac{n}{\sigma^2} \\
\frac{d^2}{d\mu d(\log\sigma)}\log p(\mu, \log\sigma|y) &= -2n\frac{\bar{y}-\mu}{\sigma^2} \\
\frac{d^2}{d(\log\sigma)^2}\log p(\mu, \log\sigma|y) &= -\frac{2}{\sigma^2}((n-1)s^2 + n(\bar{y}-\mu)^2).
\end{aligned}$$

$$\Rightarrow I(\hat{\theta}) = \begin{pmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & 2n \end{pmatrix}$$

Normal approximation:

$$p(\mu, \log\sigma|y) \approx \mathrm{N}\left(\begin{pmatrix} \mu \\ \log\sigma \end{pmatrix} \middle| \begin{pmatrix} \bar{y} \\ \log\hat{\sigma} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 1/(2n) \end{pmatrix}\right)$$

Advantages of Normal Approximation

- Application:
  - Summarizing posterior distributions by point estimate and standard error, benchmark for interpreting posterior density relative to its mode (e.g. contour plots)
  - Data reduction and summary statistics (idea of sufficient statistics)
  - Useful for debugging a computer program or checking a more elaborate method for approximating the posterior distribution
- Remark:
  - More accurate normal approximations in lower-dimensional parameter space (joint normal $\Rightarrow$ marginal normal, but marginal normal $\not\Rightarrow$ joint normal)
  - In many cases, normal approximation can be dramatically improved by transformation of parameters. The closeness of the approximation for finite n can vary substantially under different parameters.

Asymptotics

- Introduction
- Examples
- Advantages
- Large Sample Theory
- Bayesian vs Frequentist

Consistency

Taylor series expansion of $\log p(\theta|y)$ centered at the posterior mode $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \cdots$$

How does it change with the increase of sample size n?

A sequence of repeated experiments:

$y_1, \ldots, y_n \sim f(y)$ ----> 'true' underlying distribution

$p(y|\theta)$ ----> model distribution

$p(\theta)$ ----> prior distribution

Two possible scenarios

- the true data distribution is included in the parametric family

  $f(y) = p(y|\theta_0)$ for some θ_0

- the true distribution is **not** included in the parametric family

  there is no longer a true value θ_0

a value θ_0 that makes the model distribution, $p(y|\theta)$, closest to the true distribution, $f(y)$

How to measure closeness?

Consistency

- Kullback-Leibler information:

$$H(\theta) = E_f\left[\log\left(\frac{f(y)}{p(y|\theta)}\right)\right]$$
$$= \int \log\left(\frac{f(y)}{p(y|\theta)}\right) f(y)dy$$

- The KL information can be thought of as a measure of distance between the distributions $f(y)$ and $p(y|\theta)$.
- Let's assume that θ_0 is the unique minimizer of $H(\theta)$.
- Apparently, if $f(y) = p(y|\theta_0)$, then $H(\theta)$ is minimized at θ_0.
- In the following, we denote θ_0 as the minimizer of $H(\theta)$.

Consistency

- **Theorem. [Convergence in discrete parameter space]** *If the parameter space* Θ *is finite and* $P[\theta = \theta_0] > 0$. *then*

$$P[\theta = \theta_0 | y] \to 1 \quad as \ n \to \infty$$

- **Proof.** Consider the log posterior odds

$$\log\left(\frac{p(\theta|y)}{p(\theta_0|y)}\right) = \log\left(\frac{p(\theta)}{p(\theta_0)}\right) + \sum_{i=1}^{n} \log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right)$$

The last term is the sum of n *i.i.d.* r.v.s where θ and θ_0 are fixed and y_i is random with distributions f. Then

$$E_f\left[\log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right)\right] = H(\theta_0) - H(\theta) \leq 0$$

Consistency

$$\log\left(\frac{p(\theta|y)}{p(\theta_0|y)}\right) = \log\left(\frac{p(\theta)}{p(\theta_0)}\right) + \sum_{i=1}^{n} \log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right)$$

- Thus if $\theta \neq \theta_0$, the last term is the sum of n *i.i.d.* r.v.s with negative mean, which must diverge to $-\infty$ as $n \to \infty$.
- As long as $p(\theta_0) > 0$ (making the first term finite), the log posterior odds $\to -\infty$ as $n \to \infty$.
- Thus

$$\frac{p(\theta|y)}{p(\theta_0|y)} \to 0$$

  which implies $p(\theta|y) \to 0$.
- As all the probability must add to one, this implies $p(\theta_0|y) \to 1$

Consistency

- **Theorem. [Convergence in continuous parameter space]** *If θ is defined on a compact set (i.e. closed and bounded) and A is a neighborhood of θ_0 (i.e. and open set containing θ_0) with prior probability satisfying $P[\theta \in A] > 0$, then*

$$P[\theta \in A|y] \to 1 \quad \text{as} \quad n \to \infty$$

- **Proof.** See Appendix B. However the idea behind their proof is based on the idea of the discrete parameter space case.

- Note that for many problems we have discussed, Θ is not a compact set (e.g. Normal mean - $\mu \in (-\infty, \infty)$). For most problems, the compact space assumption can be relaxed. Th compact assumption is needed for the proof so Θ can be covered by a finite number of open sets and so an analogue to the discrete case can be used.

- Also note the discrete case can often be extended to allow for a infinite sample space Θ.

Consistency

Taylor series expansion of $\log p(\theta|y)$ centered at the posterior mode $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \cdots$$

A sequence of repeated experiments:

$y_1, \ldots, y_n \sim f(y)$ ----> 'true' underlying distribution

$p(y|\theta)$ ----> model distribution

$p(\theta)$ ----> prior distribution

Consistency

under regularity conditions,

$\hat{\theta} \to \theta_0$ when $n \to \infty$

Two possible scenarios

- the true data distribution is included in the parametric family

  $f(y) = p(y|\theta_0)$ for some θ_0

- the true distribution is **not** included in the parametric family

  there is no longer a true value θ_0

a value θ_0 that makes the model distribution, $p(y|\theta)$, closest to the true distribution, $f(y)$

Asymptotic Normality

Why choose $\hat{\theta}$ instead of other values?

Taylor series expansion of $\log p(\theta|y)$ centered at the posterior mode $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta-\hat{\theta})^T \left[\frac{d^2}{d\theta^2}\log p(\theta|y)\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta}) + \cdots$$

Likelihood dominating the prior distribution!

prior constant

data likelihood increases with n

$$\left[\frac{d^2}{d\theta^2}\log p(\theta|y)\right]_{\theta=\hat{\theta}} = \left[\frac{d^2}{d\theta^2}\log p(\theta)\right]_{\theta=\hat{\theta}} + \sum_{i=1}^{n}\left[\frac{d^2}{d\theta^2}\log p(y_i|\theta)\right]_{\theta=\hat{\theta}}$$

$$n\cdot\frac{1}{n}\sum_{i=1}^{n}\left[\frac{d^2}{d\theta^2}logP(y_i|\theta)\right]_{\{\theta=\hat{\theta}\}} \approx n\cdot E_f\left[\frac{d^2}{d\theta^2}logP(Y|\theta)\right]_{\{\theta=\theta_0\}} = -n\cdot J(\theta_0)$$

Consistency
under regularity conditions,
$\hat{\theta} \to \theta_0$ when $n \to \infty$

Asymptotic normality
$\sqrt{n}(\theta-\theta_0)|y \to N(0, [J(\theta_0)]^{-1})$ when $n \to \infty$

Fisher information
(if $f(y) = p(y|\theta_0)$)

Asymptotics

- Introduction
- Examples
- Advantages
- Large Sample Theory
- Bayesian vs Frequentist

Asymptotic Theory: Bayesian vs Frequentist

- Bayesian asymptotic theory

random — fixed, can replaced by θ_0

$$[I(\hat{\theta})]^{1/2}(\theta - \hat{\theta}) \mid y \sim \mathrm{N}(0, I),$$

posterior mode

Credible interval vs Confidence interval

- Frequentist asymptotic theory

fixed, can replaced by θ_0 — random

$$[I(\hat{\theta})]^{1/2}(\theta - \hat{\theta}) \mid \theta \sim \mathrm{N}(0, I)$$

maximum likelihood estimate

Asymptotics
-Counter Examples

IS THE ASYMPTOTIC THEORY APPLICABLE IN ALL SCENARIOS?

Counter Examples

- Under identified models and nonidentified parameters

- A model is called unidentified, given data y, if the likelihood, $p(y|\theta)$ is equal for a range of values of θ. The other case is exhibited by the following example:

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

- Assume that for all observations, only u or v is observed (every pair has missing data). Then

$$p(\rho|y) \propto p(\rho) \prod_{i=1}^{m} \phi(u_i) \prod_{i=1}^{m} \phi(v_j)$$

  which only depends on ρ through the prior. This is an example of a nonidentified parameter, one which has no information supplied by the data.

- For both of these problems, better data collection or information about the parameters is needed. For the example, you need to make sure you have observations where both components are not missing.

Counter Examples

- Aliasing (a special case of under identified models)

- In this case, different sets of parameter values will give the same likelihood. This is commonly seen in mixture models. For example the mixture data model

$$p(y_i|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\lambda)=\lambda\frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_1^2}(y_i-\mu_1)^2\right)+(1-\lambda)\frac{1}{\sigma_2\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_2^2}(y_i-\mu_2)^2\right)$$

- In this case if μ_1,σ_1^2 is switched with μ_2,σ_2^2 and λ is replaced by $1-\lambda$, you get the same likelihood for any $\{y_i\}$.

- The posterior distribution in this case is a (50%, 50%) mixture of two distributions that are mirror images of each other. This can’ t be normal (since it is bimodal) and can' t converge to a single point.
- The usual solution to this problem is to reparameterize the problem so the duplication disappears. For example, for this mixture model, restricting $\mu_1 \le \mu_2$ solves the problem.
- However this can get difficult in multidimentional problems. How should we order vectors? You might do something like $\|\mu_1\| \le \|\mu_2\|$.

Counter Examples

- The number of parameters increasing with sample size

- Underlying the proofs of the theorems is that the amount of information about each of the parameters increases as n increases. If this doesn't occur, consistency and asymptotic normality can't occur. For example, consider the model

$$y_i|\pi_i \sim B(n_i, \pi_i), \qquad \pi_i \sim p(\pi_i)$$

- For this model $p(\pi_i|y) \propto p(\pi_i)\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}$

  regardless of how many observations are taken. Additional observations give no further information about π_i. Only information about addition parameters π_j is collected.

- This posterior will not converge to a point. In order to converge, additional Bernoulli trials under this π_i would be needed.

Counter Examples

- Improper posterior distributions

- Implicit in these asymptotic results is that the posterior distribution is proper. For example, the consistency proof with a discrete parameter used the fact that

$$\sum_{i=1}^{k} p(\theta_k|y) = 1$$

- The solution to this problem is easy. If there is a problem, use a proper prior, which must give a proper posterior

- Note, that if there is an improper posterior, the likelihood is probably badly behaved and a likelihood analysis will also breakdown

Counter Examples

- Prior distribution excluding point of convergence

- If $p(\theta) = 0$ in the prior, $p(\theta|y) = 0$ as well. Thus if $p(\theta_0) = 0$, the posterior can't converge to θ_0, but instead will instead converge to a nearby point where $p(\theta) > 0$ (assuming it converges at all).

- To solve this problem, force the prior to satisfy $p(\theta) > 0$ for any remotely plausible θ.

Counter Examples

- Tails of the distribution

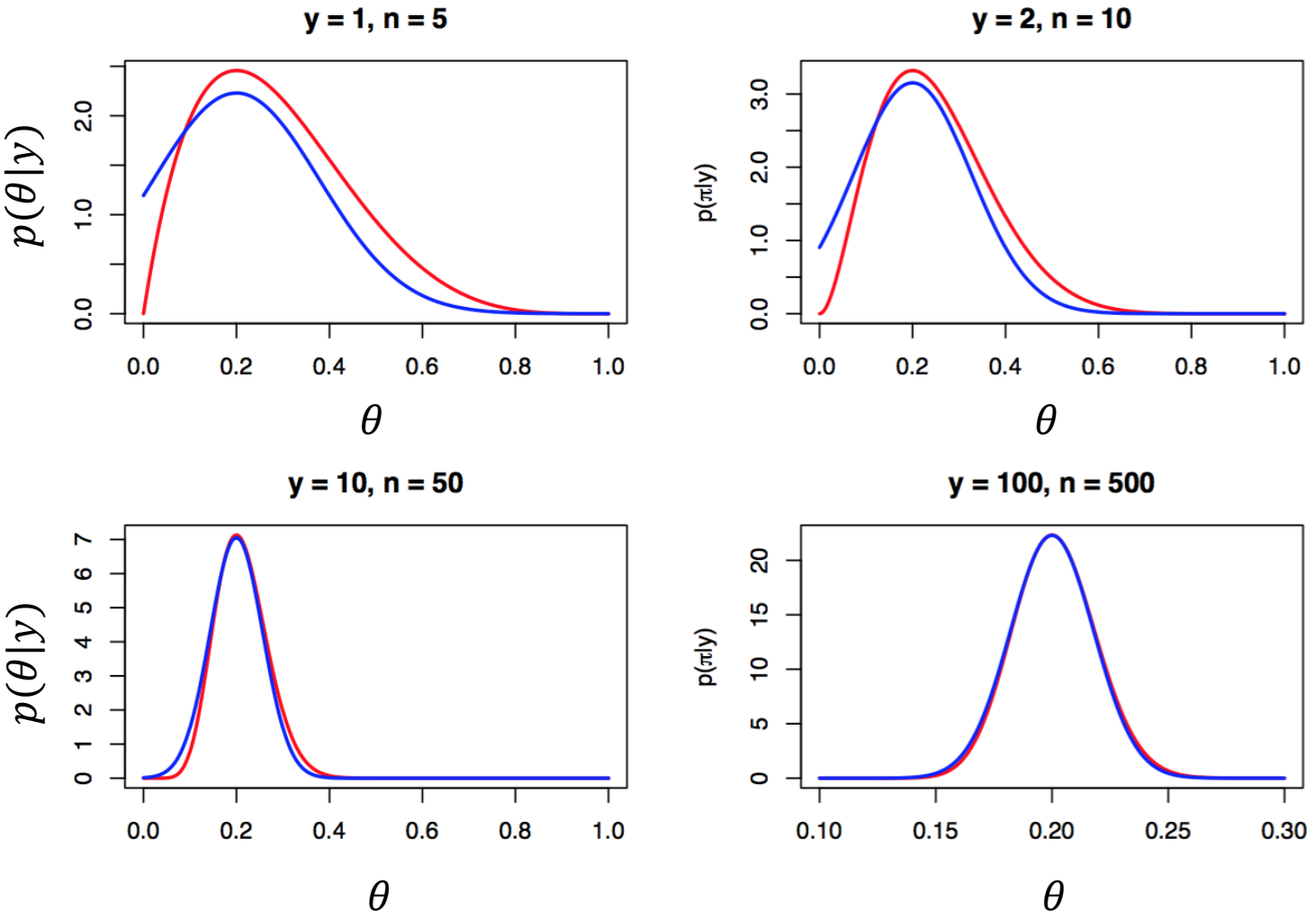
- The asymptotic normality is essentially a result about the form of the distribution in the center of its distribution. It is based a Taylor series expansion around the posterior mode (which is usually close to the posterior mean or median). It is not a result about what occurs in the tails. The normal distribution has the property that

$$p(\theta) \propto e^{-c\theta^2}$$

- However some distributions have much heavier tails (e.g. Cauchy ($p(\theta) \propto \frac{1}{\theta^2}$), Laplace ($p(\theta) \propto e^{-c|\theta|}$)), so using a normal distribution can do a bad job in the tails.

Counter Examples

- Another problem, which may be problem with finite sample sizes, is that the normal distribution takes values over an infinite range. In many problems, (e.g. binomial success probabilities), the range of the parameter is finite. However as n increases, this problem will usually disappear, as θ_0 will get further from the boundary of the parameter space.



Summary

Key Points for Today

- Simple simulation methods
  - ✓ Joint distribution, marginal distribution
  - ✓ function of parameters
- Asymptotic theory for Bayesian:
  - ✓ Key: likelihood dominating prior as sample size increase
  - ✓ Consistency of posterior mode
  - ✓ Normal approximation
  - ✓ Comparison with asymptotic theory for frequentist
  - ✓ Applicable situations illustrated by counter examples