Large Sample Inference

Statistics 220

Spring 2005

Copyright ©2005 by Mark E. Irwin

Normal Approximation to the Joint Posterior

Let $\hat{\theta}$ be the posterior mode (the maximizer of $p(\theta|y)$. Then

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

which looks like

$$c - \frac{1}{2}(\theta - \hat{\theta})^T \Sigma^{-1} (\theta - \hat{\theta}) + \ldots$$

So if the cubic and higher order terms are negligible, this is like the log of a normal density (which should occur is the posterior distribution is unimodal and roughly symmetric). Thus

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

where $I(\theta)$ is the observed information

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

So we get a similar result to that for the MLE, that is its approximately normally distributed.

Example: Binomial success probability

Lets assume the conjugate prior $\pi \sim Beta(\alpha, \beta)$. Then the posterior is

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

The derivatives are

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

which gives

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

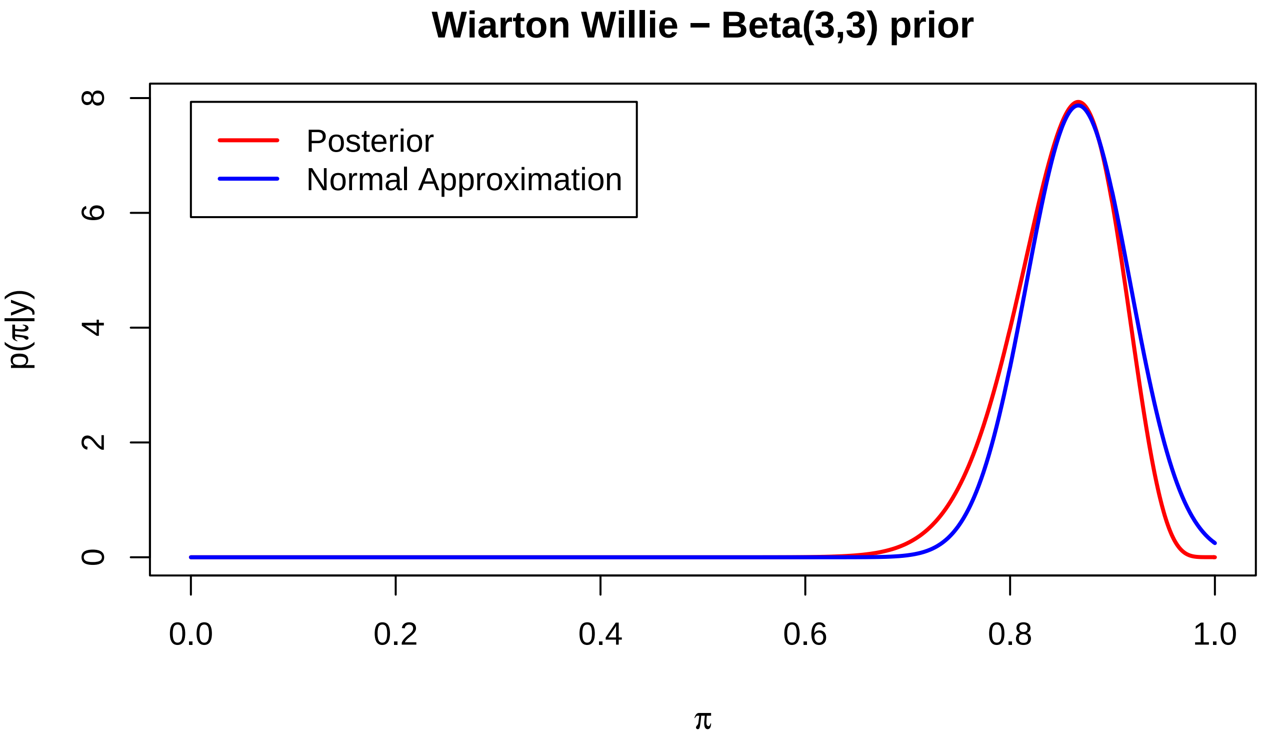
so

$$
\pi|y \overset{approx.}{\sim} N\left(\hat{\pi}, \frac{\hat{\pi}(1 - \hat{\pi})}{\alpha + \beta + n - 2}\right)
$$

For the Wiarton Willie example $(n = 41, y = 37)$ and a $Beta(3,3)$ prior

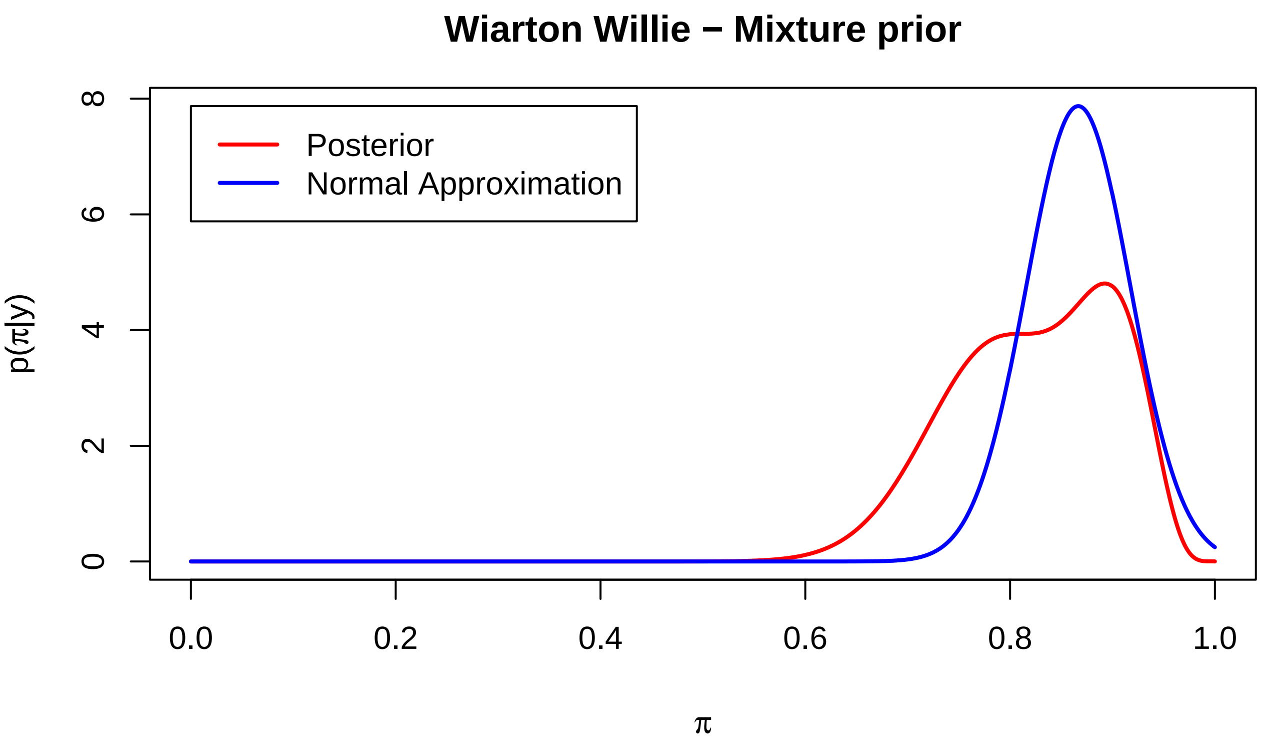
$$\hat{\pi} = \frac{3 + 37 - 1}{6 + 41 - 2} = \frac{39}{45} = 0.8667$$

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



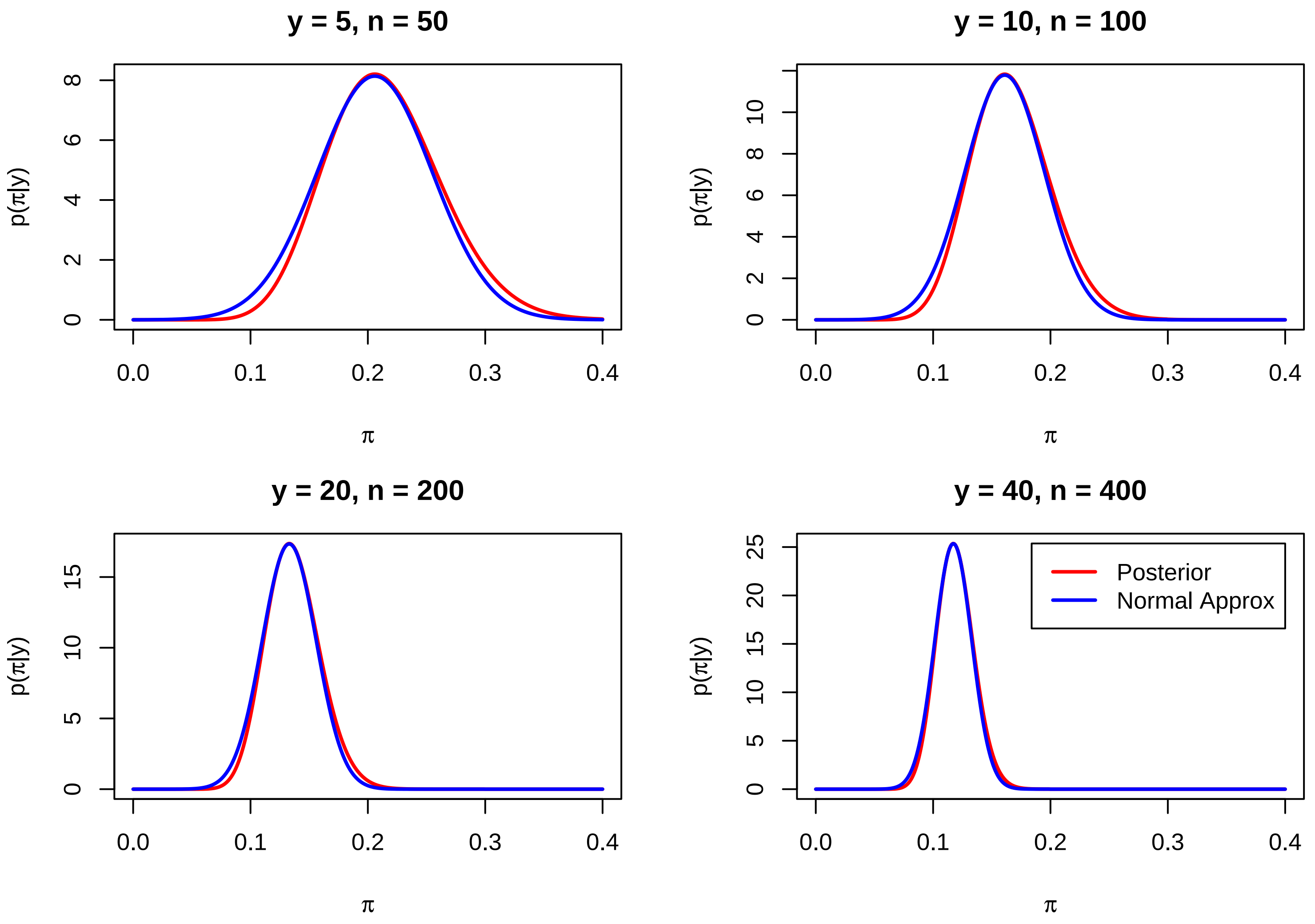
Now with the $\frac{1}{2}Beta(8,2) + \frac{1}{2}Beta(2,8)$ Mixture prior

$$\hat{\pi} = 0.8980 \qquad I(\hat{p}) = 490.11 \qquad [I(\hat{p})]^{-1/2} = 0.0452$$



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

Prior: $Beta(10, 10)$



Let $\phi = f(\theta)$ where $f(\cdot)$ is a continuous, differentiable transformation. Then both $p(\theta|y)$ and $p(\phi|y)$ both approach normal distributions. However how well they do for finite n can be different.

For example, in the discussion in the text on a normal data with unknown mean and variance, parameterizing in term of $\log \sigma$ leads to a better normal approximation.

Also, while the result talks about the parameters jointly, often the normality is closer when dealing with a subset of the parameters.

One place where this normal approximation is useful is in deriving approximate credibility intervals. In the univariate case

$$\hat{\theta} \pm z_{\alpha/2}[I(\hat{\theta})]^{-1/2}$$

is an approximate $100(1-\alpha)\%$ central credibility interval.

Do we need to worry about the normalizing constant?

In the earlier example, I worked with $c \times p(\theta|y)$ instead of $p(\theta|y)$. Note that this doesn't make a difference since

$$\begin{aligned}
\log(c \times p(\theta|y)) &= \log c + \log p(\theta|y) \\
\frac{d \log(c \times p(\theta|y))}{d\theta} &= \frac{d \log p(\theta|y)}{d\theta} \\
\frac{d^2 \log(c \times p(\theta|y))}{d\theta^2} &= \frac{d^2 \log p(\theta|y)}{d\theta^2}
\end{aligned}$$

Justification of the Asymptotic Normality and Consistency

The earlier result is imprecise, not indicating how good the normal approximation might be. Lets make it more precise.

Notation:

- $f(y)$: true distribution of the data. Assume that $y_1, \ldots, y_n \overset{iid}{\sim} f(y)$.
- $p(y|\theta)$: model distribution for the data.
- $p(\theta)$: prior distribution for the parameters

The first thing to note is that we may not be modelling the data correctly. In this case, things will converge, but not the way we might want it to be. However if $f(y) = p(y|\theta_0)$ for some θ_0, then things work the way we would like.

Kullback-Leibler information:

$$
\begin{aligned}
H(\theta) &= E\left[\log\left(\frac{f(y)}{p(y|\theta)}\right)\right] \\
&= \int \log\left(\frac{f(y)}{p(y|\theta)}\right) f(y)dy
\end{aligned}
$$

The KL information can be thought of as a measure of distance between the distributions $f(y)$ and $p(y|\theta)$. Lets assume that θ_0 is the unique minimizer of $H(\theta)$. If $f(y) = p(y|\theta_0)$, then $H(\theta)$ is minimized at θ.

For all that follows, θ_0 is the minimizer of $H(\theta)$.

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

$$P[\theta = \theta_0 | y] \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

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 iid RV where θ and θ_0 are fixed and y_i is random with distributions f. Then

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

Thus if $\theta \neq \theta_0$, the second term is the sum of n iid RVs with negative mean, which must diverge to $-\infty$ as $n \rightarrow \infty$. As long as $p(\theta_0) > 0$ (making the first term finite), the log posterior odds $\rightarrow -\infty$ as $n \rightarrow \infty$. Thus

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

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

□

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

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

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. The 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 that the discrete case can often be extended to allow for a infinite sample space Θ.

Theorem. [Asymptotic Normality of $p(\theta|y)$] *Under some regularity conditions (notably that θ_0 is not on the boundary of Θ), as $n \to \infty$*

$$\sqrt{n}(y - \theta_0) \xrightarrow{\mathcal{D}} N(0, J(\theta_0)^{-1})$$

where

$$J(\theta) = E\left[\left(\frac{d\log p(y|\theta)}{d\theta}\right)^2|\theta\right] = -E\left[\frac{d^2\log p(y|\theta)}{d\theta^2}|\theta\right]$$

Proof. Again see Appendix B, but a couple of comments.

$$\begin{aligned} I(\theta) &= -\frac{d^2}{d\theta^2}\log p(\theta|y) \\ &= -\frac{d^2}{d\theta^2}\log p(\theta) - \frac{d^2}{d\theta^2}\log p(y|\theta) \\ &= -\frac{d^2}{d\theta^2}\log p(\theta) - \sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log p(y_i|\theta) \end{aligned}$$

$$E[I(\theta)] = -\frac{d^2}{d\theta^2}\log p(\theta) + nJ(\theta)$$

so as n gets big $I(\theta) \approx nJ(\theta)$ □

Bayes vs Likelihood

Let $\hat{\theta}$ be the posterior mode and $\tilde{\theta}$ be the MLE. Under regularity conditions

- Consistency:

$$\theta | y \xrightarrow{P} \theta_0 \qquad \tilde{\theta} \xrightarrow{P} \theta_0$$

- Asymptotic normality:

$$\sqrt{n}(\theta - \theta_0)|y \xrightarrow{\mathcal{D}} N(0, J(\theta_0)) \qquad \sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N(0, J(\theta_0))$$

and

$$I(\theta_0)^{-1/2}(\theta - \theta_0)|y \xrightarrow{\mathcal{D}} N(0, I) \qquad I(\tilde{\theta})^{-1/2}(\tilde{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N(0, I)$$

So for large sample sizes, Bayes and Likelihood give equivalent answers.

Counterexamples

The results presented have assumptions behind them, and if the assumptions aren't satisfied, the consistency and asymptotic normality can break down.

- Underidentified models and nonindentified parameters

  A model is called underidentified, 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:

  $$\binom{u}{v} \sim N\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array}\right)\right)$$

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

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

which only depends on ρ through the prior. This is an example of a nonindentified 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.

- 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

$$\begin{aligned} y_i|\pi_i &\sim B(n_i, \pi_i) \\ \pi_i &\sim p(\pi_i) \end{aligned}$$

  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. For that additional Bernouilli trials under this π_i would be needed.

- Aliasing

  This is a special case of underidentified parameters. 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 and 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 $\mu_1 \leq \mu_2$ solves the problem.

However this can get difficult in multidimentional problems. How should you order vectors? You might do something like $||\mu_1|| \leq ||\mu_2||$.

- Unbounded likelihooods

  If a likelihood function is unbounded, there might not be a posterior mode in the parameter space, or it might be on the boundary.

  For example, for the mixture problem above, setting $\mu_1 = y_i$ for any observation i and letting $\sigma_1^2 \rightarrow 0$ leads to the likelihood blowing up. So for this example, there are multiple modes. As the number of observations increases, the number of modes increases and thus can't converge to a single point.

  This can usually be handled by restricting the prior to avoid these problems (e.g. force $p(\sigma_1^2) = 0$ around 0).

- 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.

- 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.

- Convergence on the edge of the parameter space

  If θ_0 is on the boundary of the parameter space problems can occur. One example comes from linkage analysis, where the recombination fraction θ is a probability that must lie between 0 and 0.5, with 0.5 corresponding to two loci occurring on different chromosomes. So with a study involving traits on two chromosomes,

  $$\hat{\theta}|y \xrightarrow{\mathcal{D}} \frac{1}{2}N(0.5, \sigma^2) + \frac{1}{2}I(\hat{\theta} = 0.5)$$

  To avoid this sort of problem, $p(\theta) > 0$ for any remotely plausible θ, or in the neighbourhood of remotely plausible values can help. Though the second part of this suggestion I find somewhat problematic. Why should I put probability on events known to be impossible, assuming that the model is correct. This is often a big assumption. For example, in the linkage analysis problem, $\theta < 0.5$ holds under certain assumptions about meiosis (dealing with interference in the crossover process). If these assumptions are wrong, $\theta > 0.5$ is possible.

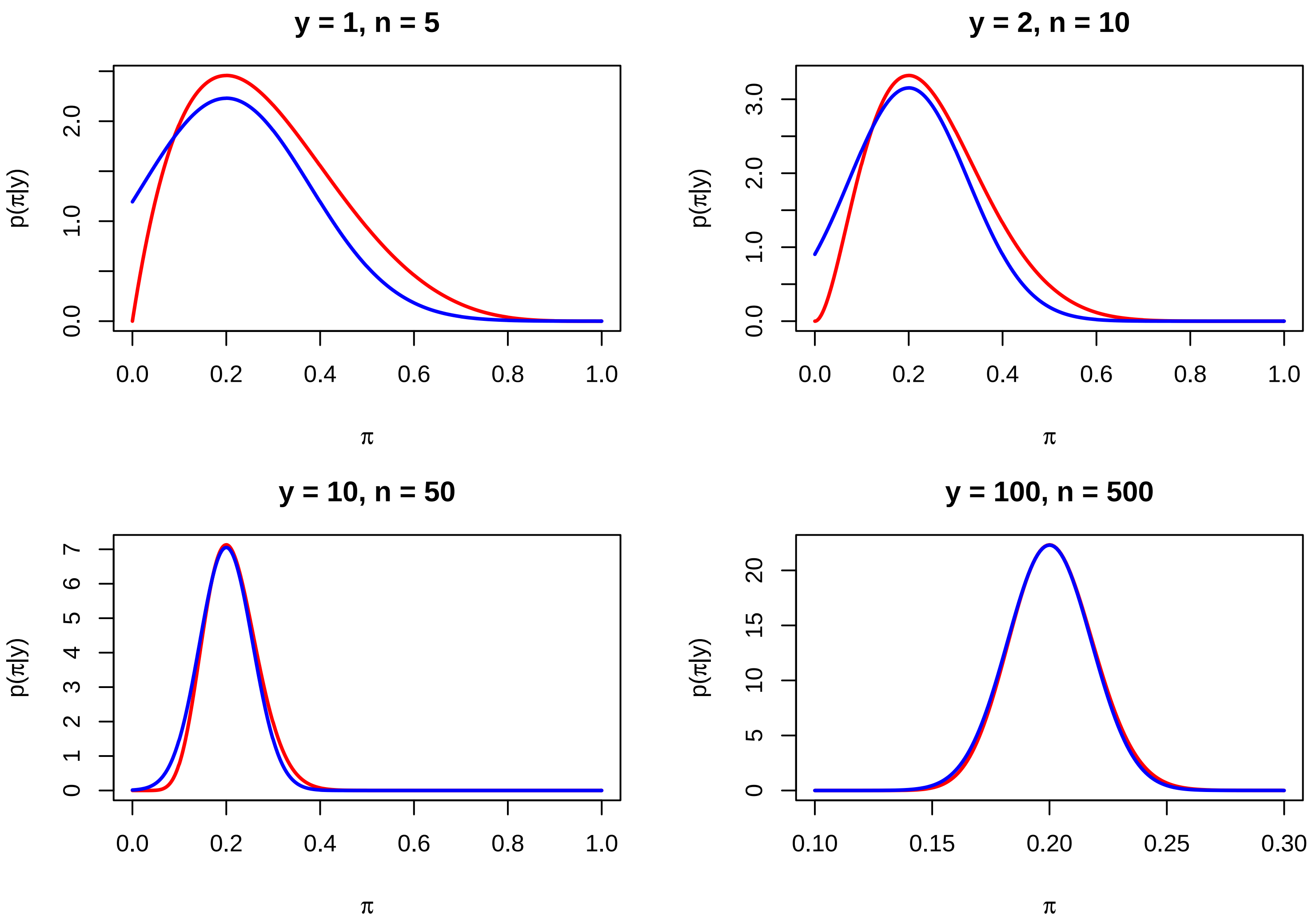
- Tails of the distribution

  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.

  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 on a standard deviation scale.


y = 1, n = 5
p(π|y)
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
π
y = 2, n = 10
p(π|y)
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
π
y = 10, n = 50
p(π|y)
0
1
2
3
4
5
6
7
0.0
0.2
0.4
0.6
0.8
1.0
π
y = 100, n = 500
p(π|y)
0
5
10
15
20
0.10
0.15
0.20
0.25
0.30
π