Prior Choice, Summarizing the Posterior

Statistics 220

Spring 2005

Informative Priors

Binomial Model:

- $y|\pi \sim Bin(n, \pi)$
- π is the success probability.
- Need prior $p(\pi)$
- Bayes used $\pi \sim U(0, 1)$
- Any density function defined on $[0, 1]$ would also be a valid prior for π.

Possible choices are

1. $Beta(a, b)$

$$p(\pi) \propto \pi^{a-1}(1-\pi)^{b-1}$$

2. Truncated $N(\mu, \sigma^2)$

$$p(\pi) \propto \frac{1}{\sigma}\phi\left(\frac{\pi-\mu}{\sigma}\right) I(0 \leq \pi \leq 1)$$

3. Mixture

$$p(\pi) = \frac{a}{2}\pi^{a-1} + \frac{b}{2}(1-\pi)^{b-1}$$

The posteriors for these priors are

1. $Beta(a, b)$

$$p(\pi|y) \propto \pi^{a-1}(1-\pi)^{b-1}\pi^{y}(1-\pi)^{n-y} = \pi^{a+y-1}(1-\pi)^{b+n-y-1}$$

So the posterior is $Beta(a+y, b+n-y)$.

2. Truncated normal

$$p(\pi|y) \propto \frac{1}{\sigma}\phi\left(\frac{\pi-\mu}{\sigma}\right)\pi^{y}(1-\pi)^{n-y}I(0 \leq \pi \leq 1)$$

This doesn't have a nice form.

3. Mixture

$$\begin{aligned} p(\pi|y) &\propto \left(\frac{a}{2}\pi^{a-1} + \frac{b}{2}(1-\pi)^{b-1}\right)\pi^y(1-\pi)^{n-y} \\ &= \frac{a}{2}\pi^{a+y-1}(1-\pi)^{n-y} + \frac{b}{2}\pi^y(1-\pi)^{b+n-y-1} \end{aligned}$$

This is a mixtures of betas.

Conjugate Priors

- The first case is an example of a conjugate prior.
- A prior is said to be **conjugate** to the sampling density if the resulting posterior is a member of the same parametric family as the prior.
- In the first case: binomial likelihood $\times$ beta prior $=$ beta posterior
- However the second example clearly isn't conjugate.
- In the third case: binomial likelihood $\times$ power law mixture prior $=$ beta mixture posterior

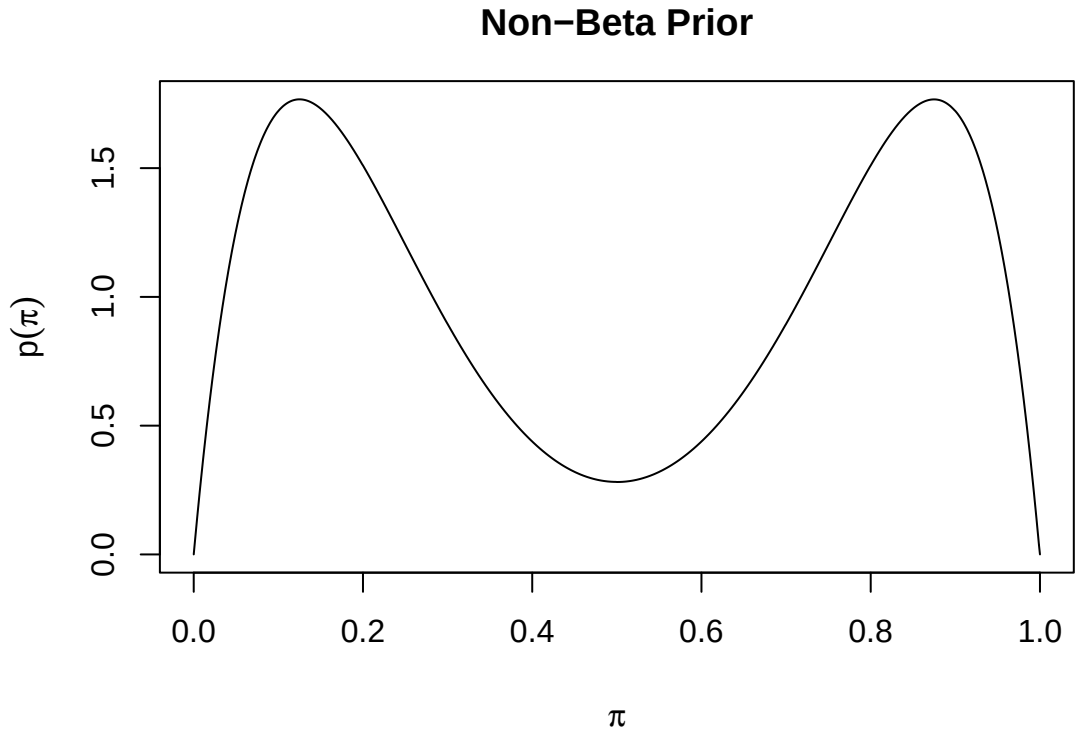
  This is almost conjugate but not quite.

- Advantages of conjugate priors:
  - Easy to deal with - mathematically and computationally
  - Interpretable as additional data

    For example, a $Beta(a, b)$ prior in the binomial success problem can be thought to be equivalent to seeing a data set earlier that had a successes and b failures.

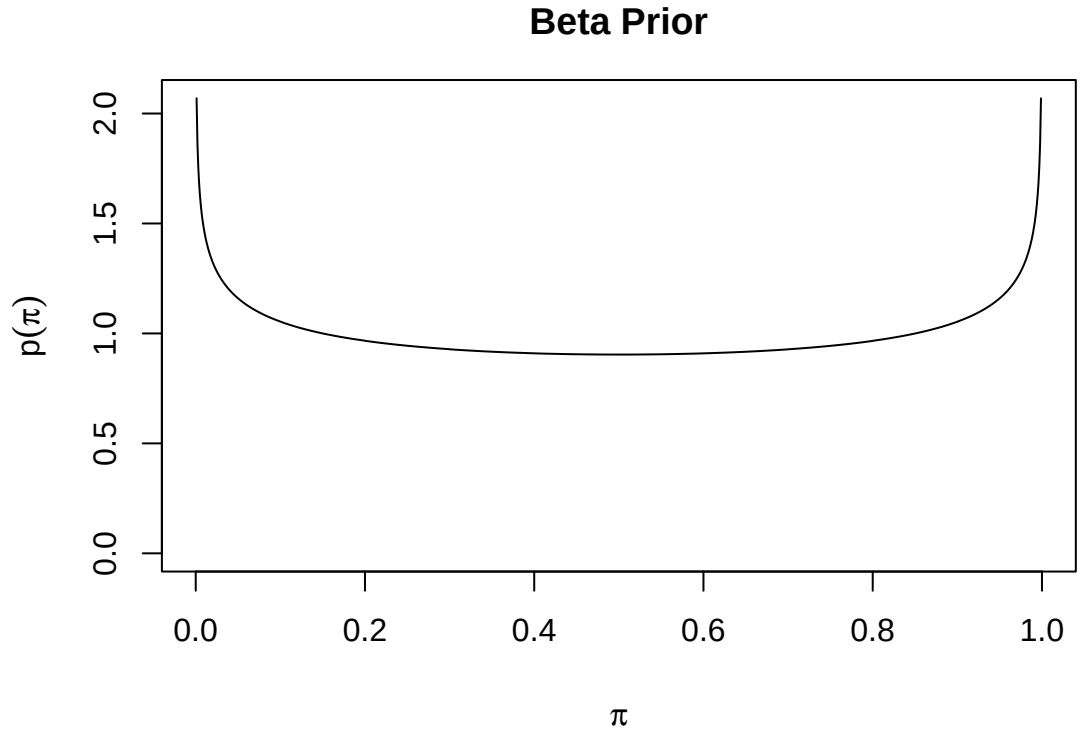
- Disadvantages of conjugate priors:
  - Can be overly restrictive
  - Some prior beliefs can not be described.

For example, suppose you wanted to use a beta prior with binomial data, but your prior beliefs matched



There is no beta distribution that looks like this.

About the closest you could get would be something that looks like



which doesn't get the bimodal nature correct.

- Whether a conjugate prior exists depends on the form of the likelihood function. Most cases do not have conjugate distributions.

- About the only case where conjugate prior are guaranteed to exist is whether the data distribution is a member of the exponential family. This includes many important cases, such as

  – Normal
  – Binomial, Multinomial
  – Poisson
  – Gamma
  – Beta

  For example, let $y|\theta \sim N(\theta, \sigma^2)$, $\theta \sim N(\mu_0, \tau_0^2)$ (assume that σ^2, μ_0, and τ_0^2 are all known quantities). Then $\theta|y \sim N(\mu_1, \tau_1^2)$ where

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

Summarizing Posterior Distributions

When doing an analysis, we need summaries of the posterior distribution.

In a univariate problem, such as for the binomial success probability or for the normal mean (with known variance) a plot of the posterior density is useful.

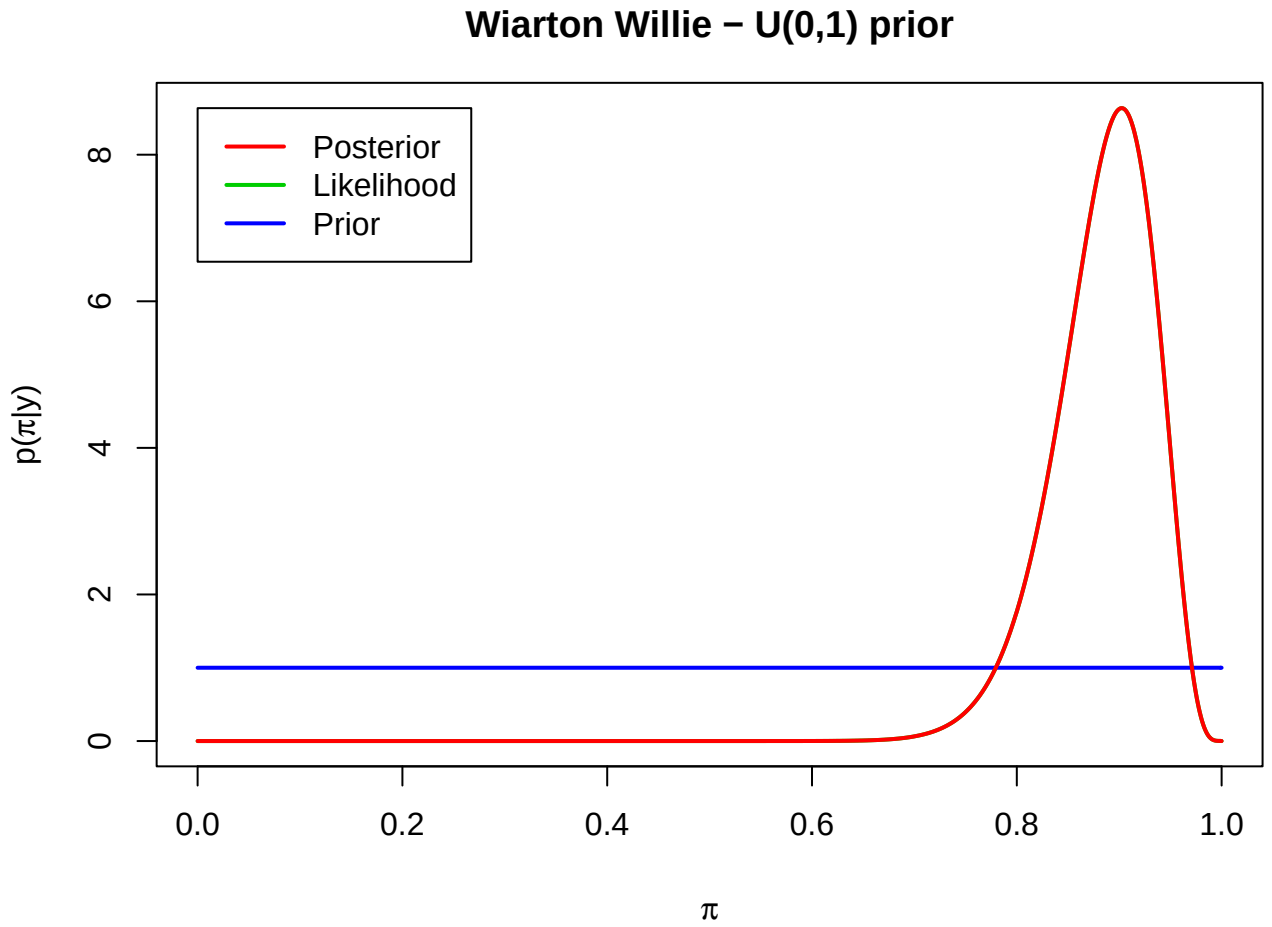
If the prior is conjugate, the plot is easy.

If the prior is non-conjugate, the plot isn't much harder.

Examples (Wiarton Willie - $n = 41, y = 37$):

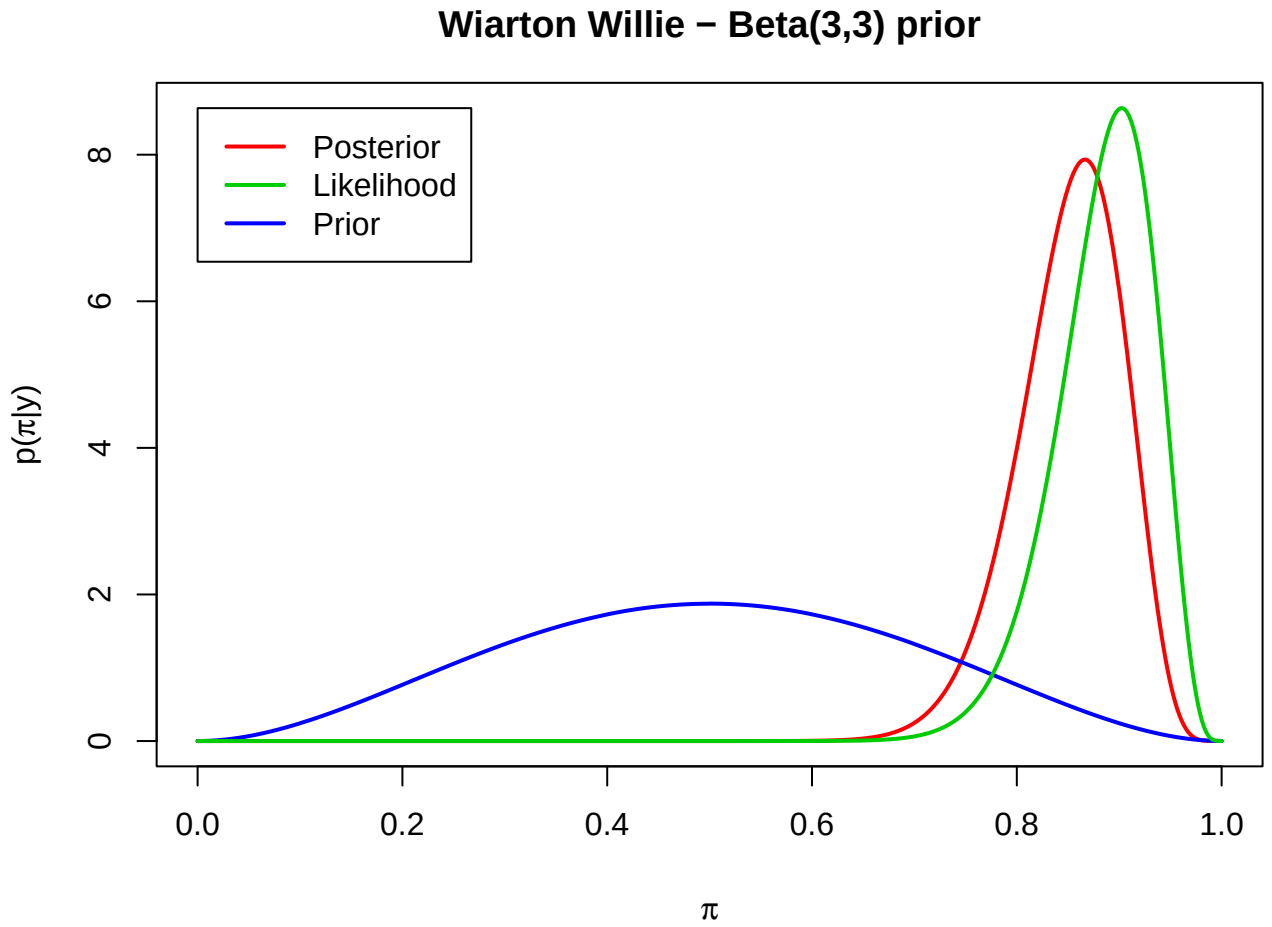
- $\pi \sim U(0,1) = Beta(1,1)$

  $\pi|y \sim Beta(38,5)$

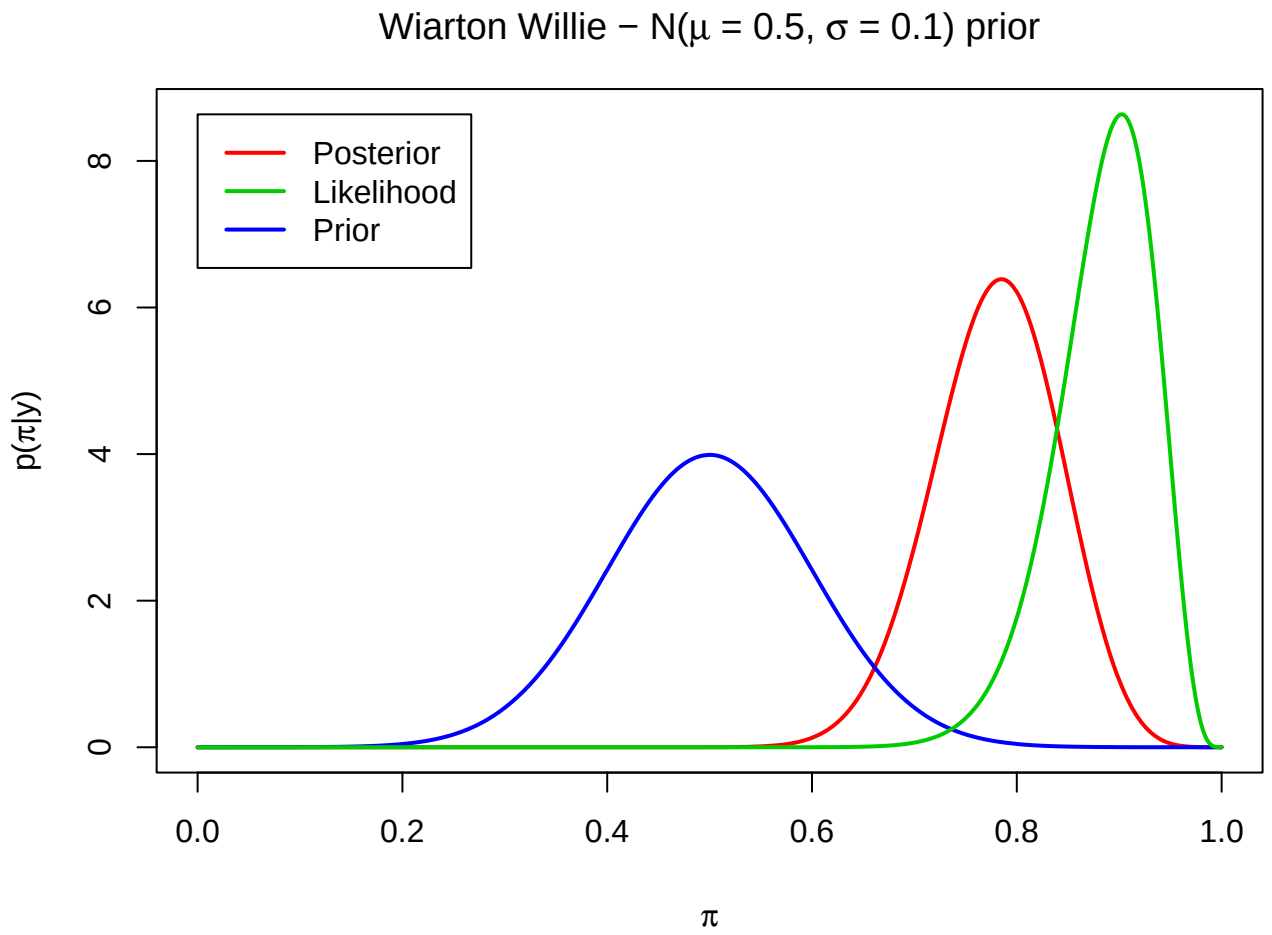


- $\pi \sim Beta(3, 3)$

  $\pi|y \sim Beta(40, 7)$

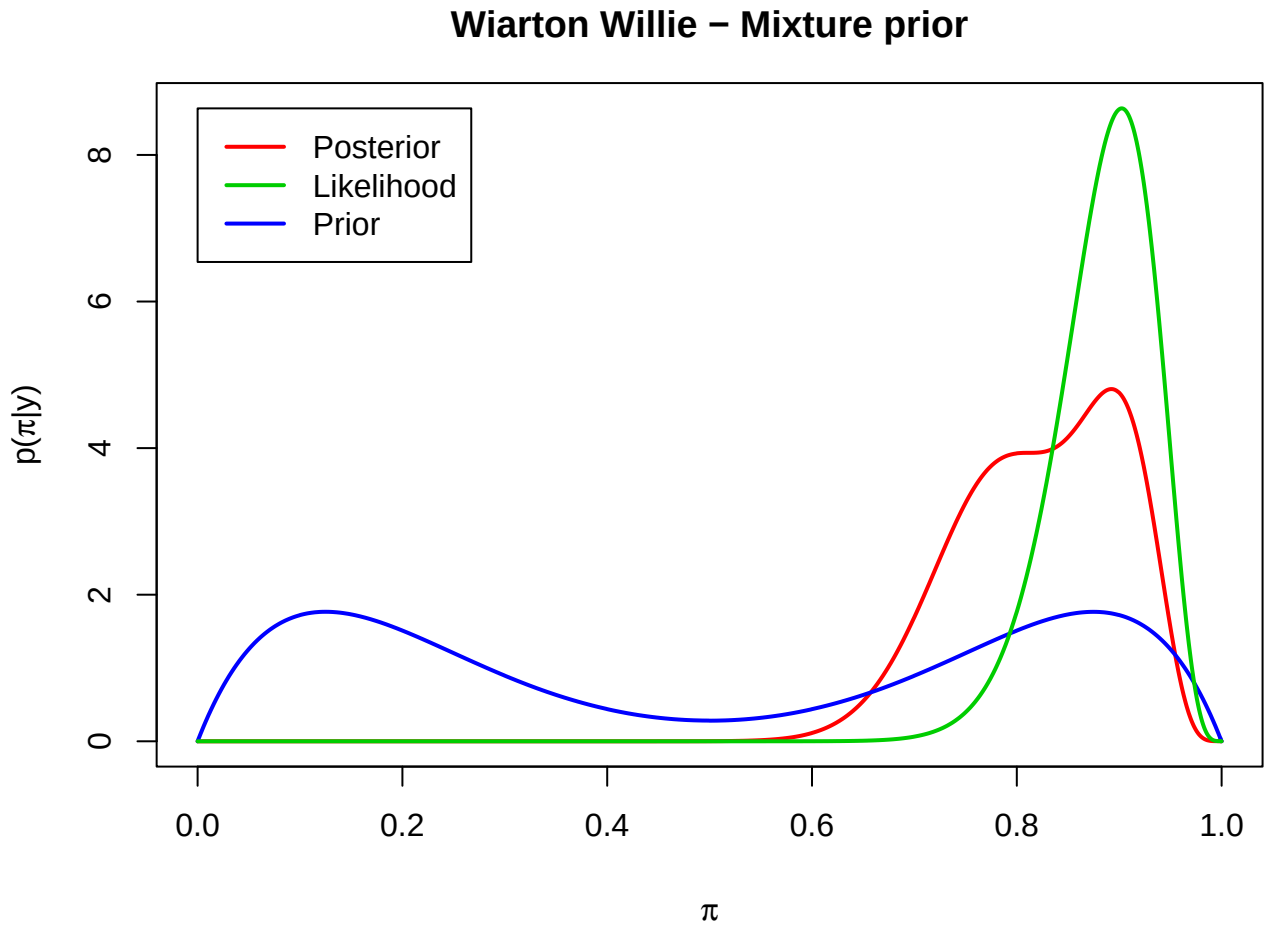


- $\pi \sim N(\mu = 0.5, \sigma^2 = 0.01)I(0 \leq \pi \leq 1)$



- Mixture prior - $\pi \sim \frac{1}{2}Beta(8,2) + \frac{1}{2}Beta(2,8)$

  $\pi|y \sim \frac{1}{2}Beta(45,6) + \frac{1}{2}Beta(39,12)$



In three of the four examples, the posterior density has a nice form. However, for the third one, the posterior density doesn't.

We can get something for plotting very easily

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}$$

but

$$p(y) = \int p(\theta)p(y|\theta)d\theta \approx \sum_{i=1}^{m} p(\theta_0 + \Delta\theta i)p(y|\theta_0 + \Delta\theta i)\Delta\theta = c$$

i.e. calculate the unnormalized density at m equally spaced points on the interval $[\theta_0, \theta_1]$ where $\Delta\theta = \frac{\theta_1 - \theta_0}{m}$

Then

$$p(\theta|y) \approx \frac{p(\theta)p(y|\theta)}{c}$$

Note that this approximation will work well when $\Delta\theta$ is small and m is large.

What is classified as a small $\Delta\theta$ and a large m depends on how smooth $p(\theta)p(y|\theta)$ is.

Of course other numerical quadrature methods (e.g. Simpson's Rule) could be used to calculate the normalization constant of the posterior.

Also this adjustment is only needed if you want compare this posterior distribution to another distribution (e.g. the prior).

If you just want to see the shape of the posterior, just plot $p(\theta)p(y|\theta)$ versus θ. The correction just relabels the y axis on the plot.

Numerical Summaries of Posterior Distributions

When summarizing a posterior distribution we are usually interested in two things, location and spread.

- Measures of location:

  There are three common measure of location used

  - Posterior Mean: $E[\theta|y]$
  - Posterior Median: $\text{Med}(\theta|y)$
  - Posterior Mode: $\arg\max_\theta p(\theta|y)$

- These choices, among being common summaries of location, can also be justified in a decision theory formulation. They are the optimal estimators under different loss functions

  - Posterior Mean: $E[\theta|y] = \arg\min_a E[(\theta - a)^2|y]$
  - Posterior Median: $\text{Med}(\theta|y) = \arg\min_a E[|\theta - a||y]$
  - Posterior Mode: $\arg\max_\theta p(\theta|y) = \arg\min_a E[I(\theta \neq a)|y]$ (0-1 loss)

- Measures of spread:

  The common choices for measuring spread are

  – $\mathrm{Var}(\theta|y)$
  – $\mathrm{SD}(\theta|y)$

  Other possibilities, though less common are

  – Posterior IQR
  – Posterior MAD (Mean Absolute Deviation): $E[|\theta - E[\theta|y]|\,|y]$

Credible Sets

It is also useful to find regions of the parameter space that accounts for most of the posterior probability.

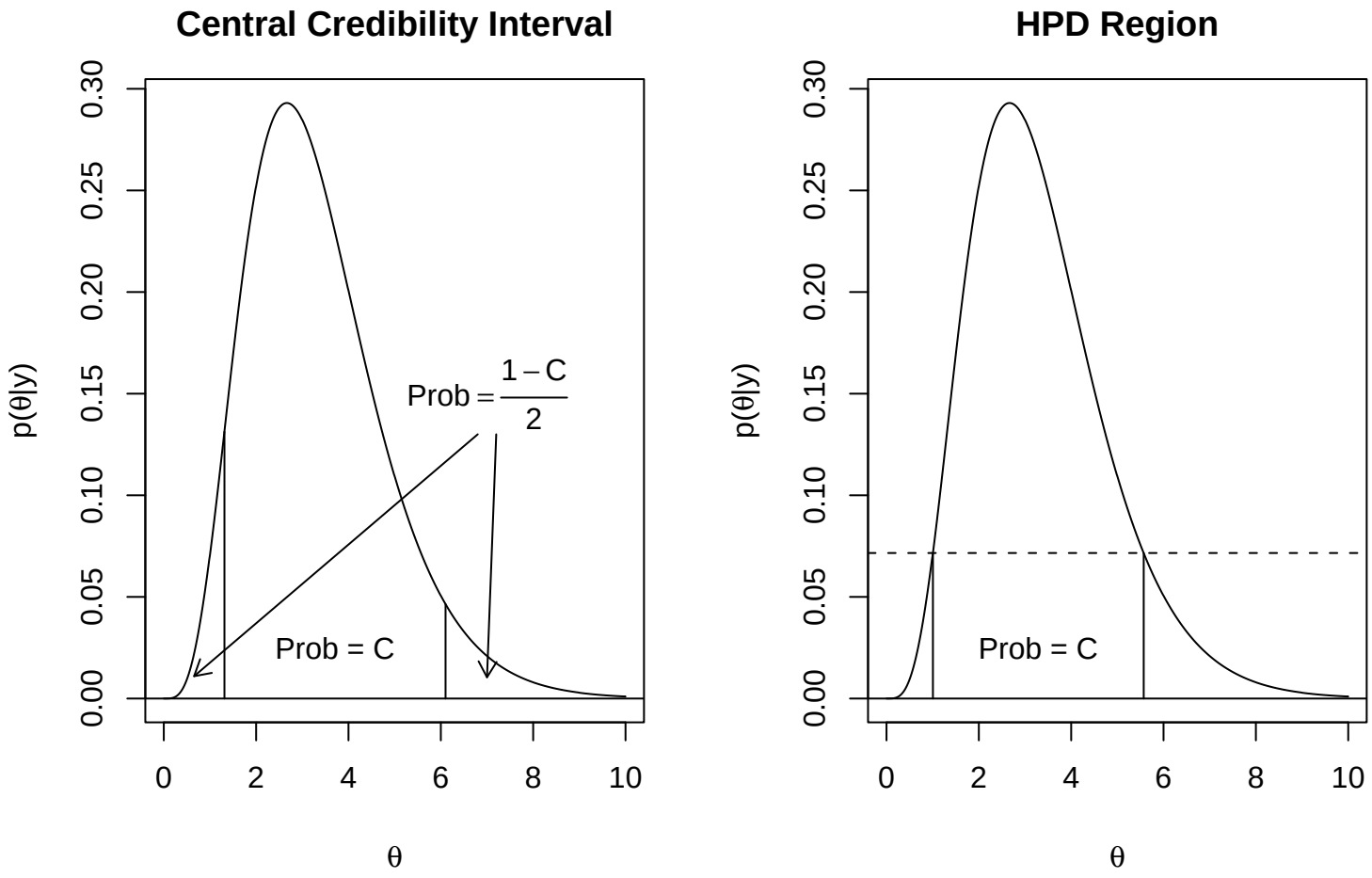
Let $A \subset \Theta$ be some subset of the parameter space for θ. Then A is a $100C\%$ credible set for θ if

$$P[\theta \in A|y] = C$$

The most common approach to credible sets are central credible intervals. These are defined as the interval $[c_l, c_u]$ such that

$$\frac{1-C}{2} = \int_{-\infty}^{c_l} p(\theta|y)d\theta \quad \text{and} \quad \frac{1-C}{2} = \int_{c_u}^{\infty} p(\theta|y)d\theta$$

An alternative is the $100C\%$ highest posterior density (HPD) region, which is defined as the smallest region of the parameter space with probability C.



Central interval: (1.313, 6.102); length = 4.789

HPD interval: (1.006, 5.571); length = 4.565

The central interval is usually easier to determine as it just involves finding quantiles of the posterior distribution.

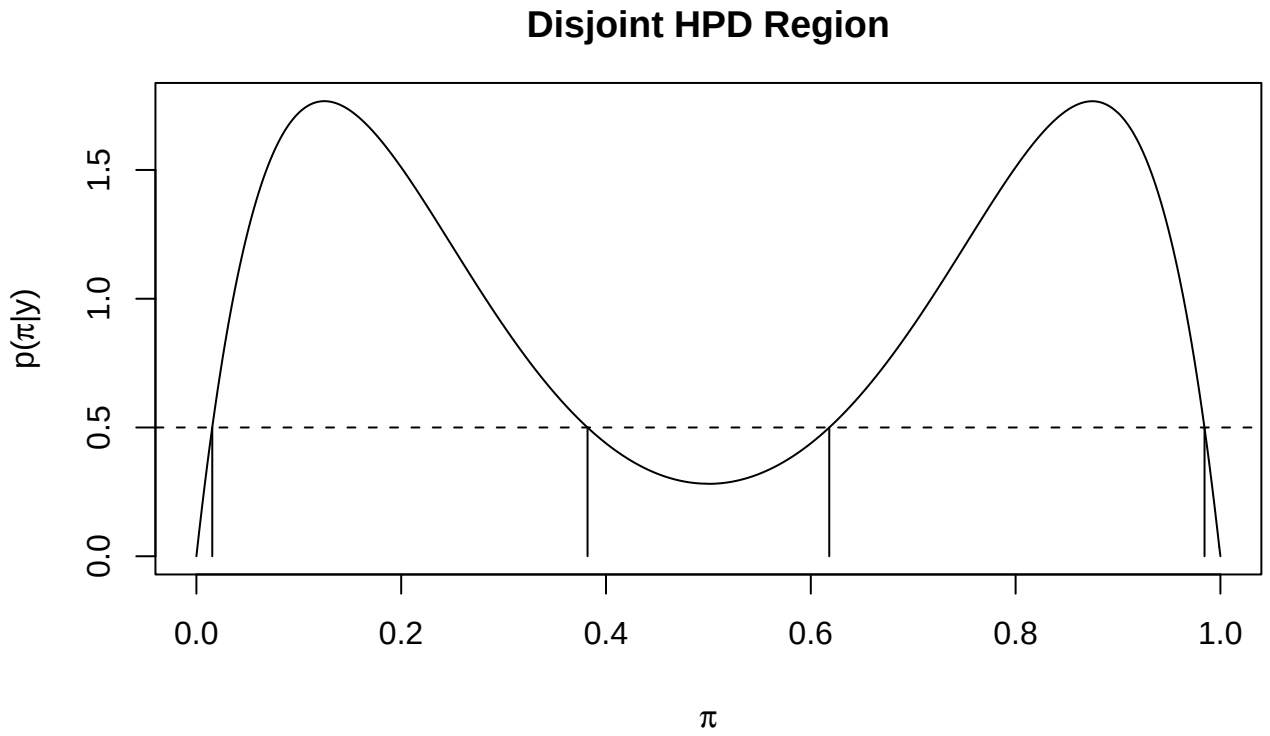
The HPD region is more difficult to calculate. It involves dealing with two main steps

1. Determining $A_c = \{\theta : p(\theta|y) \geq c\}$

2. Calculating $P[\theta \in A_c|y]$

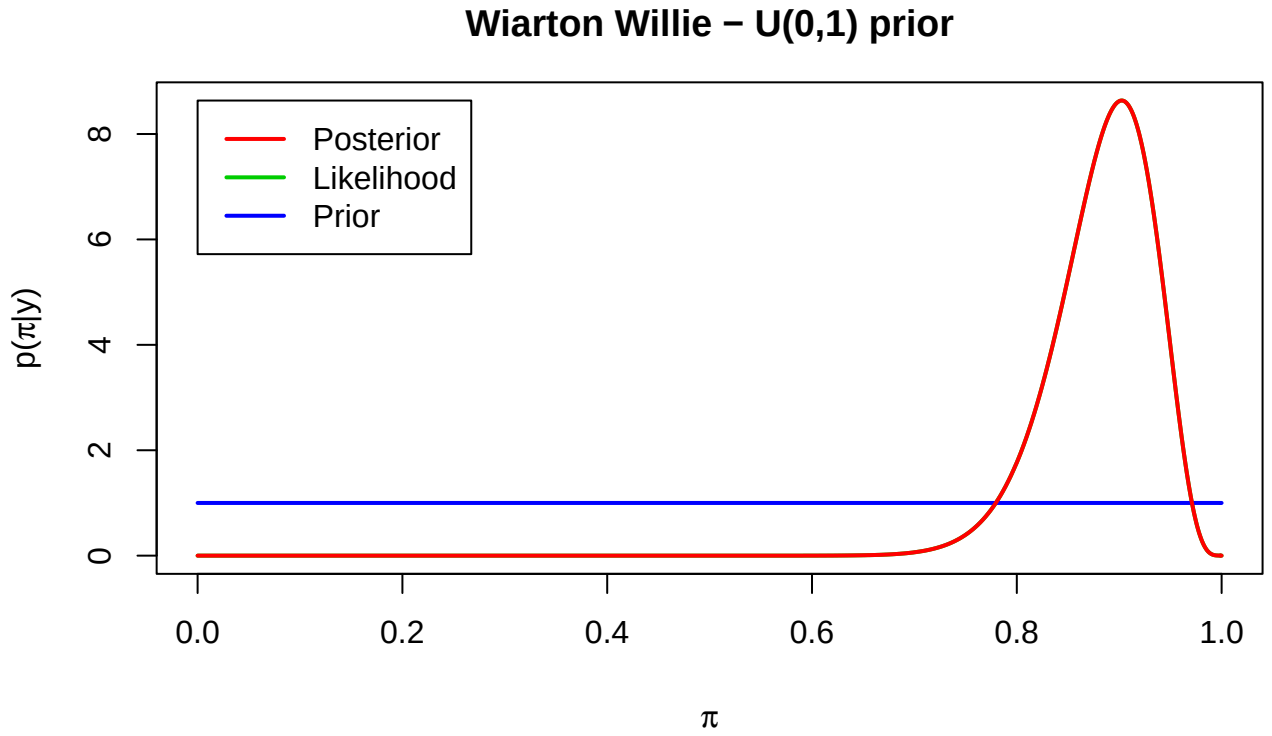
The HPD region is then given by the c where $P[\theta \in A_c|y] = C$

In the example graph, it wasn't too bad since the posterior is unimodal. However if the posterior is multimodal, A_c may be a bunch of disjoint sets



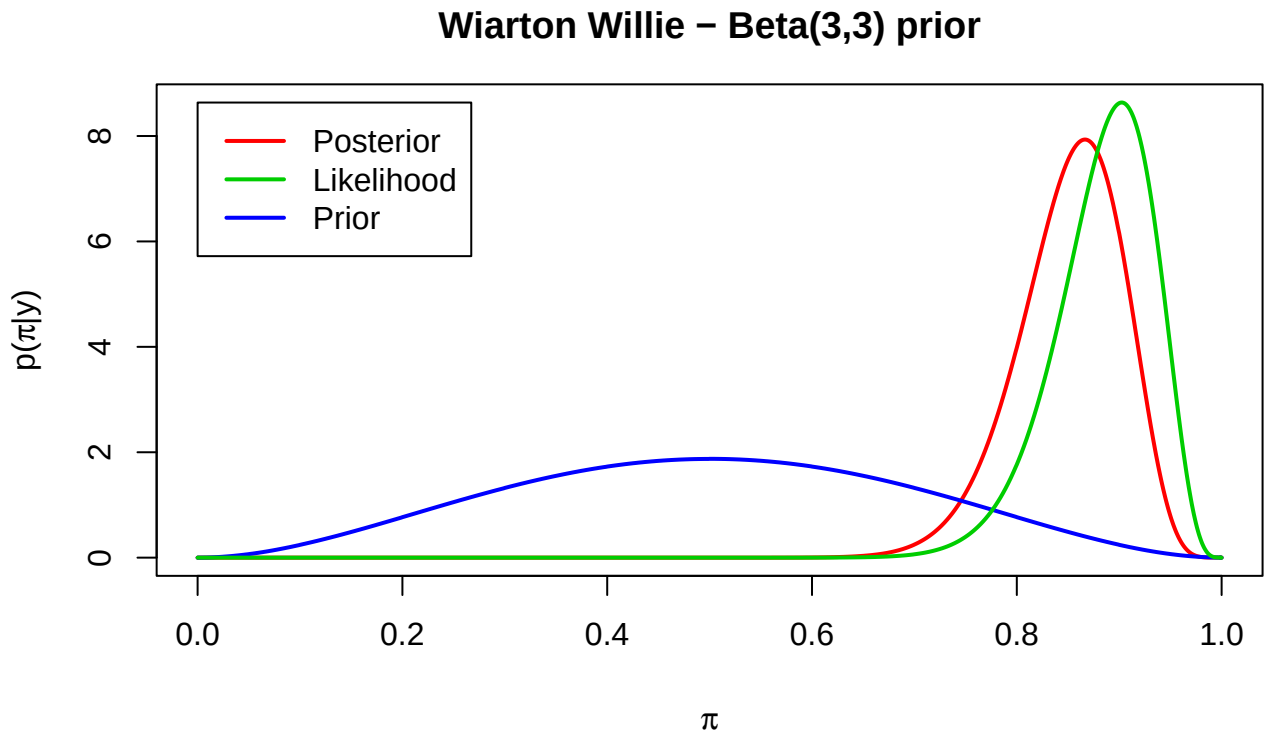
Especially in big problems, determining the number of modes is difficult so HPD regions are rarely used today.

- $\pi \sim U(0,1) = Beta(1,1) \qquad \pi|y \sim Beta(38,5)$



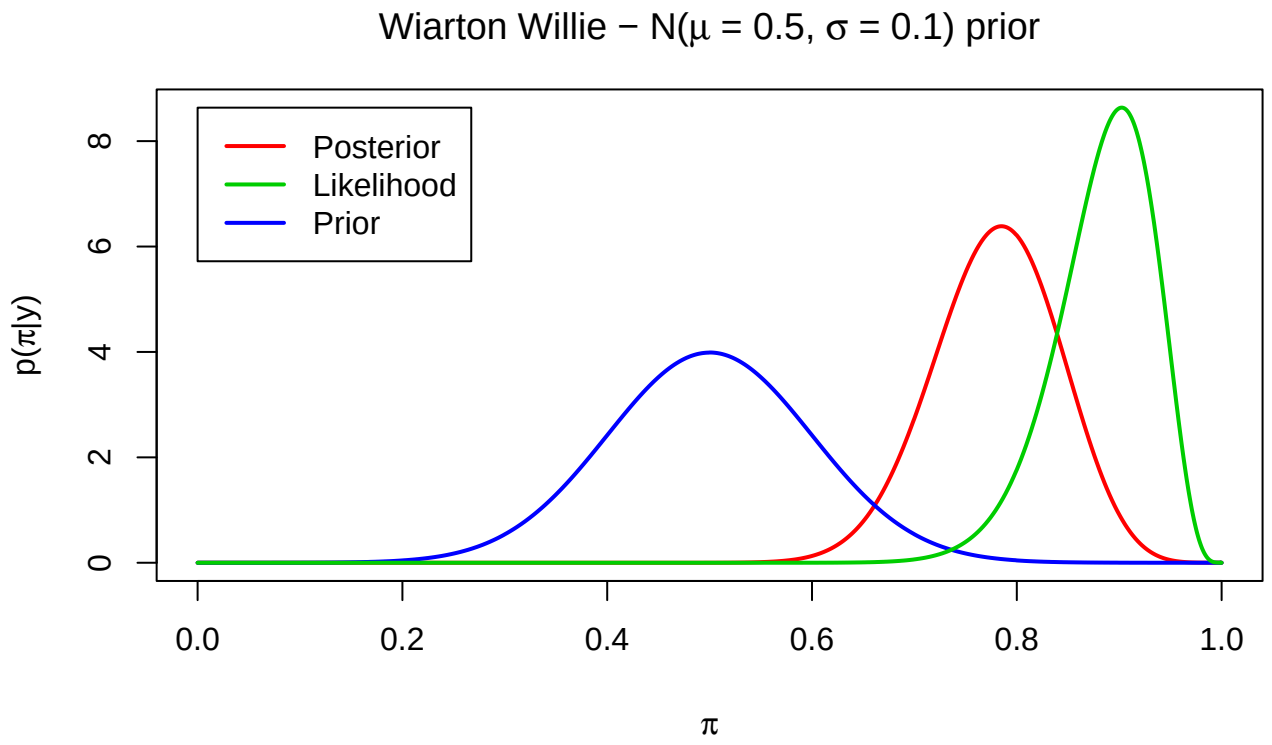
Distribution	Mean	Mode	SD	95% Central Cred. Int.
Prior	0.5	???	0.289	(0.025, 0.975)
Posterior	0.884	0.902	0.048	(0.774, 0.960)

- $\pi \sim Beta(3,3) \qquad \pi|y \sim Beta(40,7)$



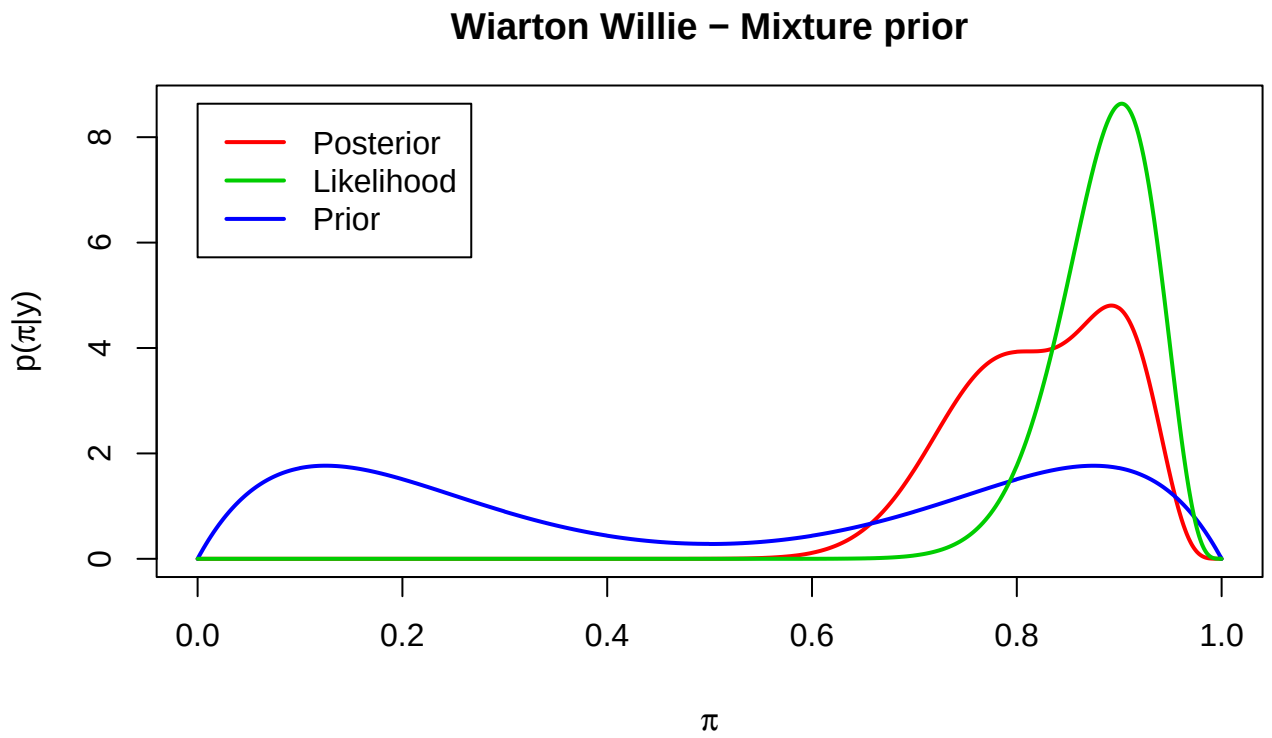
Distribution	Mean	Mode	SD	95% Central Cred. Int.
Prior	0.5	0.5	0.189	(0.147, 0.853)
Posterior	0.851	0.867	0.051	(0.737, 0.937)

- $\pi \sim N(\mu = 0.5, \sigma^2 = 0.01)I(0 \leq \pi \leq 1)$



Distribution	Mean	Mode	SD	95% Central Cred. Int.
Prior	0.5	0.5	0.1	(0.304, 0.696)
Posterior	???	???	???	(???, ???)

- $\pi \sim \frac{1}{2}Beta(8,2) + \frac{1}{2}Beta(2,8)$ $\quad\quad$ $\pi|y \sim \frac{1}{2}Beta(45,6) + \frac{1}{2}Beta(39,12)$



Distribution	Mean	Mode	SD	95% Central Cred. Int.
Prior	0.5	??? & ???	???	(???, ???)
Posterior	0.823	???	???	(???, ???)

Except for the situation where the prior is nice (e.g. Beta), there are not nice formulas for many of the summaries.

Numerical methods are needed to calculate these

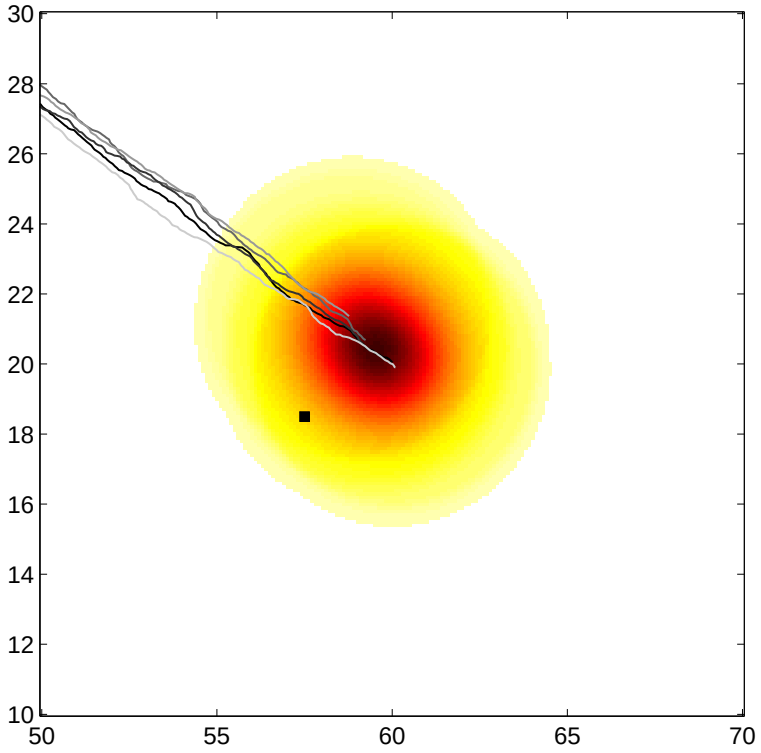
- Solving equations - Newton-Raphson, bisection, etc
- Optimization - Newton-Raphson, EM, etc
- Numerical Quadrature - Simpson's Rule, Gaussian Quadrature, etc
- Combinations of these to determine credibility intervals

There is another approach that can be used as an alternative for most of these - **Simulation**.

It can be used to find posterior means, variances, quantiles, etc for parameters, and functions of parameters.

Example: Risk estimation in Command and Control

The threat that enemy tank (at position (x_t, y_t)) poses to a target (at position (x_0, y_0)) depends on:



- the distance from the target

$$d_t = \sqrt{(x_t - x_0)^2 + (y_t - y_0)^2}$$

- the damage function $\delta(d_t)$, which is usually a non-linear decreasing function of d_t.

Calculating $E[\delta(d_t)]$, even by quadrature methods can be very difficult.

However approximating this expectation is easily done by simulation.