VectorBiTE Training 2018
Methods Workshop

Introduction to Bayesian Statistics

www.vectorbite.org

Learning Objectives

1. Understand the basic principles underlying Bayesian modeling methodology
2. Introduce how to use Bayesian inference for real-world problems
3. Introduce computation tools to perform inference for simple models in R (how to turn the Bayesian crank)
4. Appreciate the need for sensitivity analysis, model checking and comparison, and the potential dangers of Bayesian methods.

Recall: Bayes Theorem

Bayes Theorem allows us to relate the conditional probabilities of two events A and B:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

What is Bayesian Inference?

In the Bayesian approach our probabilities numerically represent rational beliefs. Bayes rule provides a rational method for updating those beliefs in light of new information and incorporating/quantifying uncertainty in those beliefs.

Thus, Bayesian inference is an approach for understanding data inductively.

What is Bayesian Inference?

We can re-write Bayes rule in terms of our parameters, θ and our data, Y:

$$\Pr(\theta|Y) = \frac{\Pr(Y|\theta)\Pr(\theta)}{\Pr(Y)}$$

The LHS is the main quantity of interest in a Bayesian analysis, the posterior, denoted $f(\theta|Y)$:

$$\overbrace{f(\theta|Y)}^{\text{Posterior}} \propto \overbrace{\mathcal{L}(\theta;Y)}^{\text{Likelihood}} \times \overbrace{f(\theta)}^{\text{Prior}}$$

Bayesian methods provide

1. models for rational, quantitative learning
2. parameter estimates with good statistical properties
3. estimators that work for small and large sample sizes
4. parsimonious descriptions of data, predictions for missing data, and forecasts for future data
5. a coherent computational framework for model estimation, selection and validation

Classical vs Bayesian

The fundamental differences between classical and Bayesian methods is what is fixed and what is random in an analysis

Paradigm	Fixed	Random
Classical	param (θ)	data (Y)
Bayesian	data (Y)	param (θ)

Why/Why Not Bayesian Statistics?

Pros

1. If $f(\theta)$ & $\mathcal{L}(\theta; Y)$ represent a rational person's beliefs, then Bayes' rule is an optimal method of updating these beliefs given new info (Cox 1946, 1961; Savage 1954; 1972).
2. Provides more intuitive answers in terms of the probability that parameters have particular values.
3. In many complicated statistical problems there are no obvious non-Bayesian inference methods.

Cons

1. It can be hard to mathematically formulate prior beliefs (choice of $f(\theta)$ often ad hoc or for computational reasons)
2. Posterior distributions can be sensitive to prior choice.
3. Analyses can be computationally costly.

Steps to Making Inference

1. Research question
2. Data collection
3. Model $Y_i \approx f(X_i)$
4. Estimate the parameter in the model with uncertainty
5. Make inference

The difference between Classical and Bayesian lies in step 4: (C) uses maximum likelihood estimate (MLE), and (B) derives a posterior distribution.

Example: Estimating the probability of a rare event

Suppose we are interested in the prevalence of an infectious disease in a small city. A small random sample of 20 individuals will be checked for infection.

- Interest is in the fraction of infected individuals

$$\theta \in \Theta = [0, 1]$$

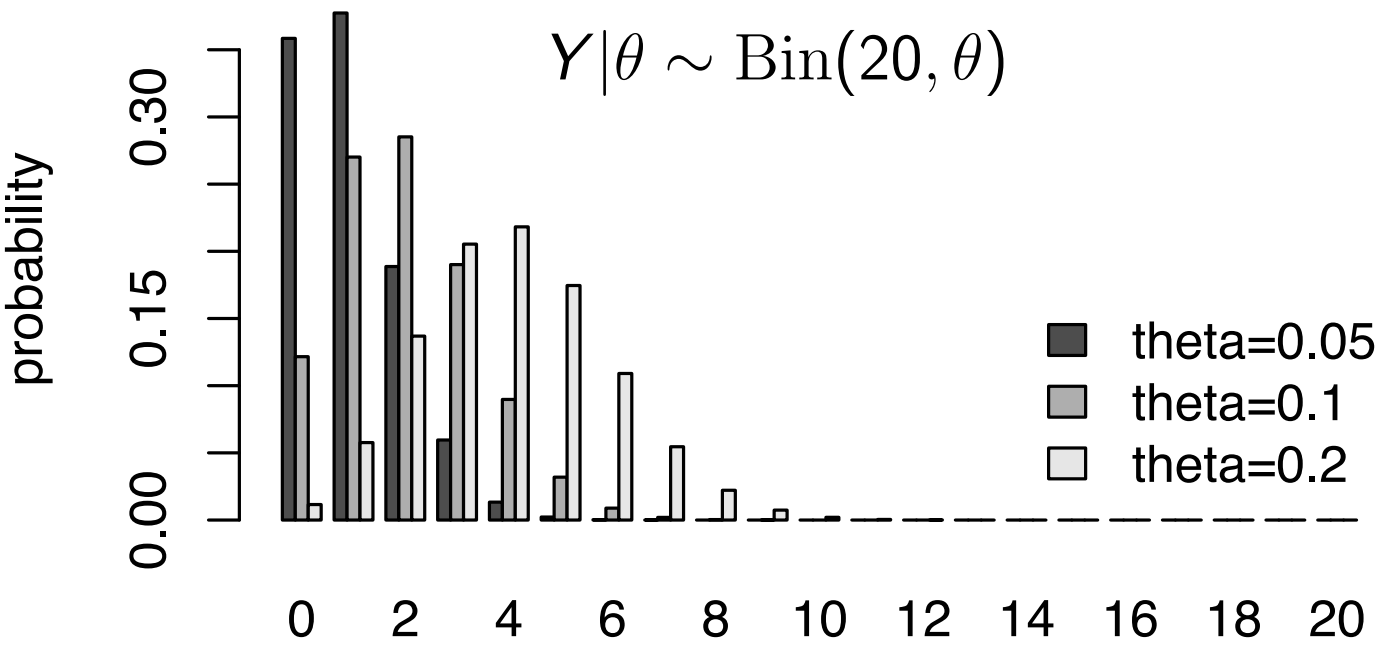
- The data records the number of infected individuals

$$y \in \mathcal{Y} = \{0, 1, \ldots, 20\}$$

Example: Likelihood/sampling model

Before the sample is obtained, the number of infected individuals is unknown.

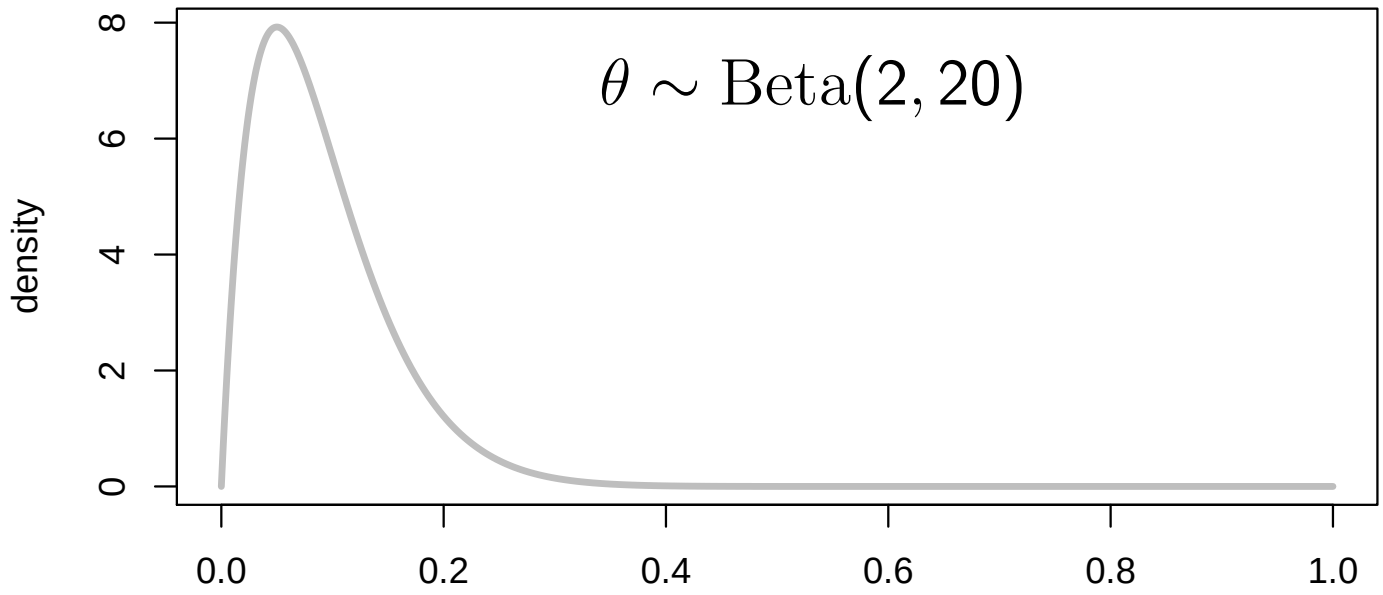
- Let Y denote this to-be-determined value
- If θ were known, a sensible sampling model is



Example: Prior

Other studies from various parts of the country indicate that the infection rate ranges from about 0.05 to 0.20, with an average prevalence of 0.1.

- Moment matching from a beta distribution (a convenient choice) gives the prior



Example: Posterior

The prior and sample model combination:

$$\theta \sim \text{Beta}(a, b)$$
$$Y|\theta \sim \text{Bin}(n, \theta)$$

and an observed y (the data), leads to the posterior

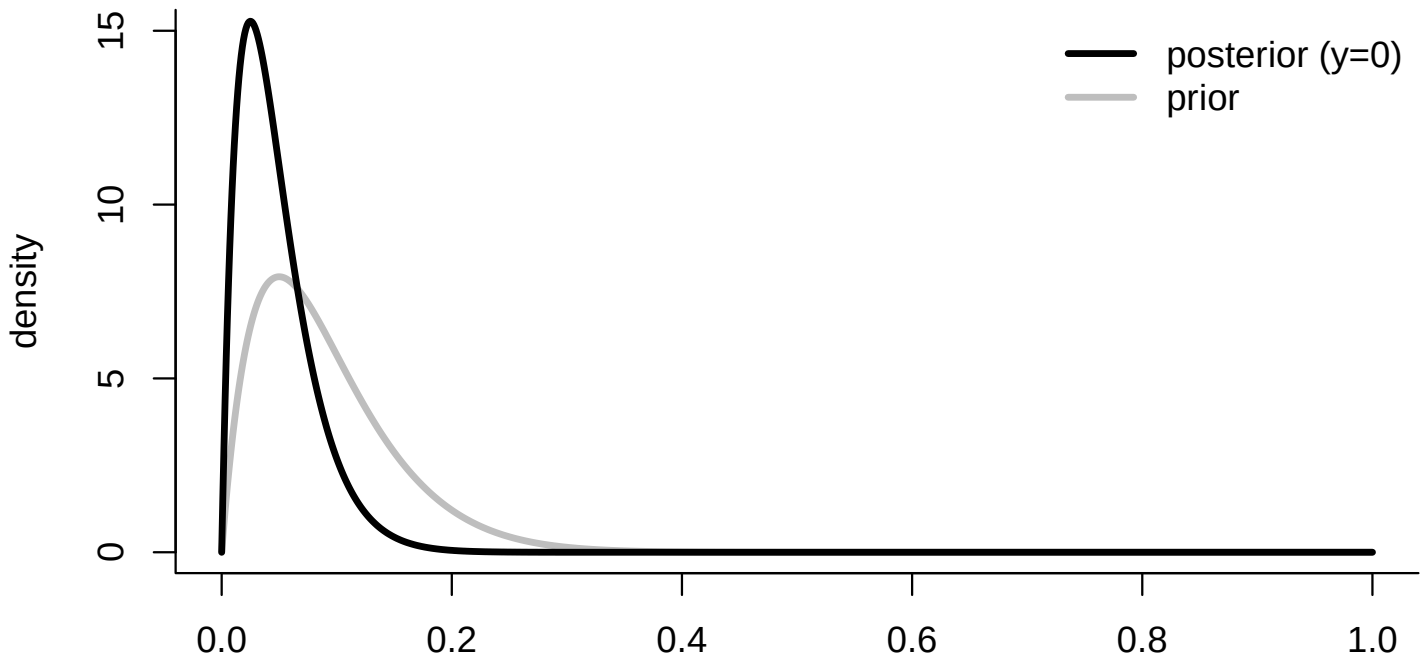
$$p(\theta|y) = \text{Beta}(a + y, b + n - y)$$

Example: Posterior

For our case, we have $a = 2$, $b = 20$, $n = 20$.

If we don't find any infections, then $y = 0$ our posterior is:

$$p(\theta|y = 0) = \text{Beta}(2, 40)$$



Example: Sensitivity Analysis

How influential is our prior?

The posterior expectation is

$$\mathrm{E}\{\theta|Y=y\} = \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0$$

a weighted average of the sample mean and the prior expectation:

$$\theta_0 = \frac{a}{a+b} \quad \rightarrow \text{ prior expectation (or guess)}$$

$$w = a+b \quad \rightarrow \text{ prior confidence}$$

Example: A non-Bayesian approach

A standard estimate of a population proportion, θ is the sample mean $\bar{y} = y/n$, the fraction of infected people in the sample.

If $y = 0$, this gives zero, so reporting the sampling uncertainty is crucial (e.g., for reporting to health officials).

The most popular 95% confidence interval for a population proportion is the Wald Interval:

$$\bar{y} \pm 1.96\sqrt{\bar{y}(1-\bar{y})/n}.$$

This has the correct *asymptotic* coverage, but $y = 0$ is still problematic!

Exercise: is a treatment for cancer effective?

We have data on n cancer patients that have been given a particular treatment. Our outcome variable is whether or not it was "effective":

Patient	Effectiveness	Numerical Data
1	N	0
2	N	0
3	Y	1
⋮	⋮	⋮

The appropriate sampling model for each patient is a Bernoilli:

$$Y_i|\theta \overset{iid}{\sim} f(Y|\theta) = \text{Bern}(\theta)$$

where θ is the success rate of the treatment. Based on this, write down the likelihood for the n patients. Then, assuming a $\text{Beta}(a, b)$ prior for θ find the posterior distribution for $\theta|Y$. Does this look familiar?

Conjugate Bayesian Models

Some sets of priors/likelihoods/posteriors exhibit a special relationship called *conjugacy*. This happens if the posterior distribution has the same form as the prior. For instance, in our above Beta-Binomial/Bernoilli examples:

$$\theta \sim \text{Beta}(a, b)$$
$$Y|\theta \sim \text{Bin}(n, \theta)$$
$$\theta|Y \sim \text{Beta}(a^*, b^*)$$

Are all posteriors in the same family as the priors? **No**
Conjugacy is a nice special property, but most of the time this isn't the case and getting an analytic form of the posterior distribution can be hard or impossible.

What do you do with a Posterior?

- Summarize important aspects of the posterior
  - mean, median, mode, variance...
- Check sensitivity of posterior to prior choice
- Say what range of parameters is consistent with the observed data given our prior information
- Make predictions

Posterior Summaries

If we are interested in point-summaries of our posterior then the (full) distribution allows us many choices. For example for the Beta-Binomial model, we found that

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

So, for example:

$$\text{E}[\theta|Y] = \frac{a + y}{a + b + n}$$

$$\text{mode}(\theta|Y) = \frac{a + y - 1}{a + b + n - 2} \quad \dagger$$

† a.k.a. the maximum *a posteriori estimator* (MAP)

Prior Sensitivity

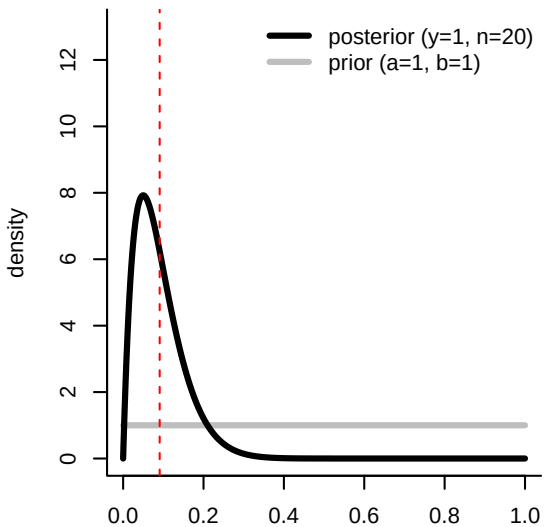
As we saw earlier for the Beta-Binomial model, the posterior expectation can be written as a weighted average of information from the prior and the data

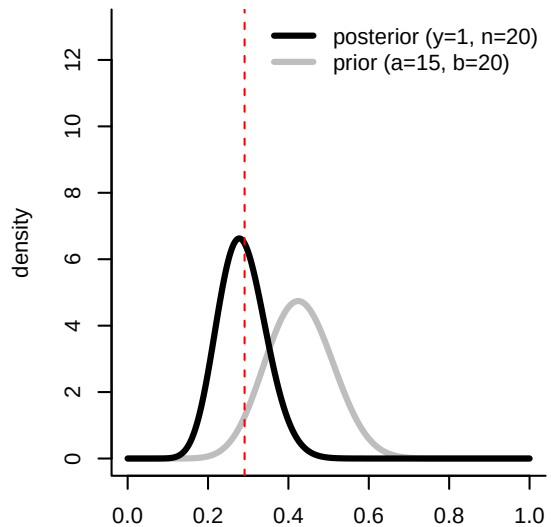
$$\mathrm{E}\{\theta|Y=y\} = \frac{n}{a+b+n}\bar{y} + \frac{a+b}{a+b+n}\theta_0.$$

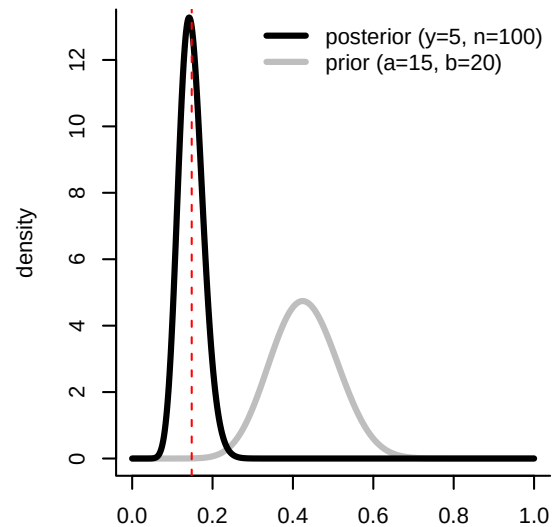
Thus a and b can be interpreted here as "prior data" where a is the number of "prior successes" and $a+b$ is the "prior sample size". When $n \gg a+b$ most of our information comes from the data instead of the prior.

Prior Sensitivity

We usually visualize the prior vs. posterior to check for sensitivity, since we don't have analytic ways, in general.







Confidence Regions

An interval $[l(y), u(y)]$, based on the observed data $Y = y$, has 95% Bayesian coverage for θ if

$$P(l(y) < \theta < u(y)|Y = y) = 0.95$$

The interpretation: it describes your information about the true value of θ after you have observed $Y = y$.

Such intervals are typically called credible intervals, to distinguish them from frequentist confidence intervals. Both are referred to as CIs.

Quantile-based (Bayesian) CI

Perhaps the easiest way to obtain a credible interval is to use the posterior quantiles.

To make a $100 \times (1-\alpha)\%$ quantile-based CI, find numbers $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ such that

1. $P(\theta < \theta_{\alpha/2} | Y = y) = \alpha/2$
2. $P(\theta > \theta_{1-\alpha/2} | Y = y) = \alpha/2$

The numbers $\theta_{\alpha/2}, \theta_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ posterior quantiles of θ.

Example: Binomial sampling + uniform prior

Suppose out of $n = 10$ conditionally independent draws of a binary random variable we observe $Y = 2$ ones (successes).

Using a uniform prior distribution (a.k.a., $\mathrm{Beta}(1, 1)$) for θ, the posterior distribution is $\theta | Y = 2$ $\mathrm{Beta}(1 + 2, 1 + 10 - 2)$.

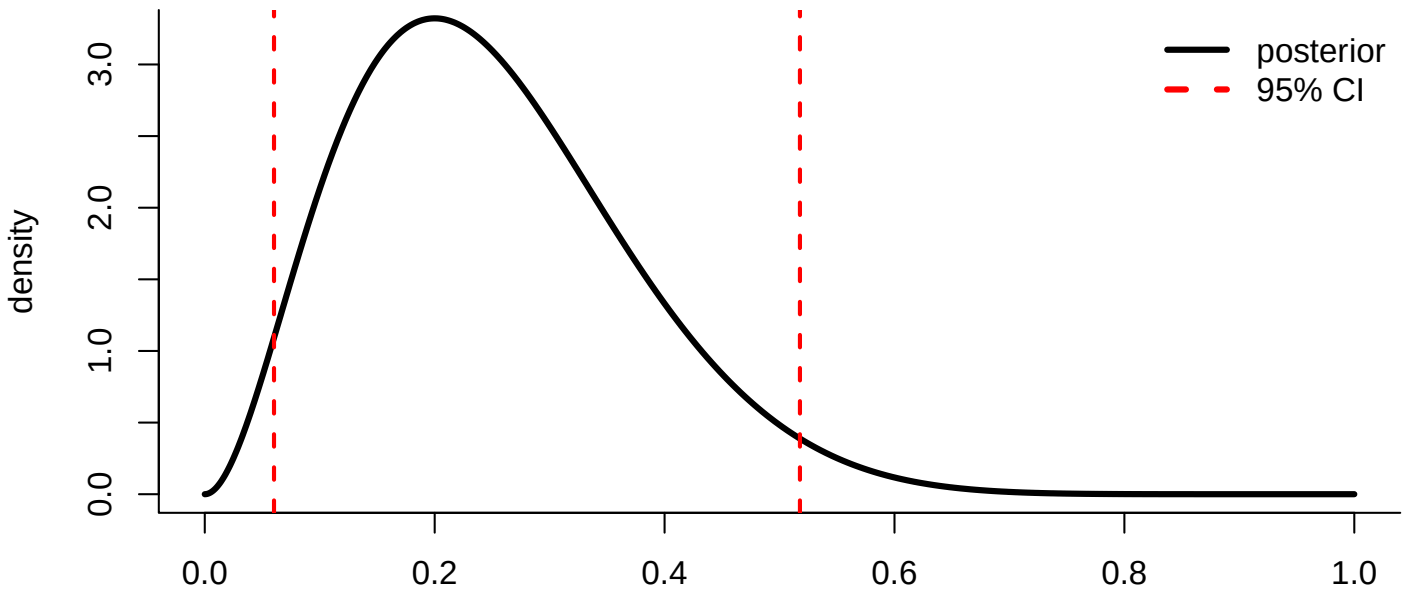
A 95% CI can be obtained from the 0.025 and 0.975 quantiles of this beta distribution (e.g., using R):

```
> round(qbeta(p=c(0.025, 0.975), 3, 9), 2)
[1] 0.06 0.52
```

Thus the posterior probability that $\theta \in [0.06, 0.52]$ is 95%.

Example: Binomial sampling + uniform prior

One drawback: Notice that there are θ-values outside the quantile-based CI that have higher probability [density] than some points inside the interval.



HPD region

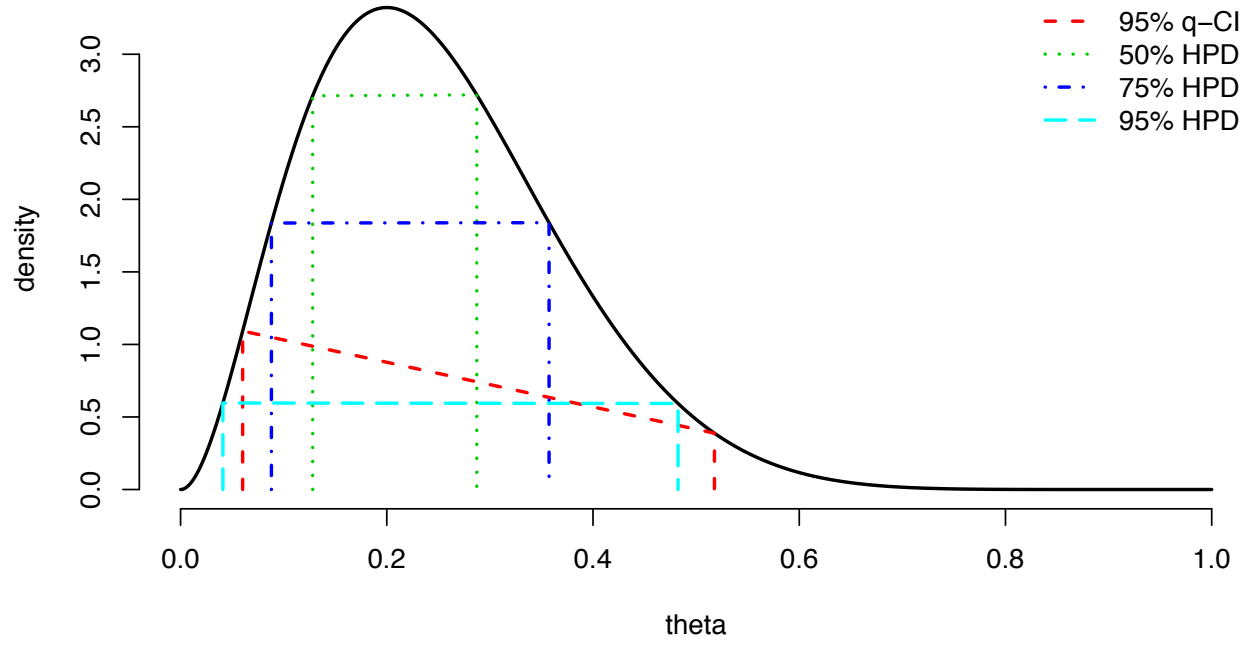
An alternative is a more restrictive type of interval. A $100 \times (1 - \alpha)\%$ highest posterior density (HPD) regions is the part of parameter space, $s(y)$, such that:

1. $P(\theta \in s(y)|Y = y) = 1 - \alpha$
2. If $\theta_a \in s(y)$ and $\theta_b \notin s(y)$ then $P(\theta_a|Y = y) > P(\theta_b|Y = y)$

So that all points in the HPD region have higher probability density than those outside.

Example: Binomial sampling + uniform prior

Numerically, we collect the highest density points with cumulative density greater that $1 - \alpha$:



The 95% HPD region is [0.04, 0.48] which is narrower than the quantile-based CI, yet both contain 95% probability.

Numerical Methods

Most of the time we can't get a nice analytic form for a posterior distribution. If we go back to the full Bayes theorem:

$$\Pr(\theta|Y) = \frac{\mathcal{L}(\theta; Y)f(\theta)}{\Pr(Y)}$$

We are usually specifying the likelihood and the prior but we often don't know the normalizing constant in the denominator. Without this, the probabilities don't properly integrate to 1 and we can't make probability statements. We need a way to approximate the distribution

Monte Carlo Simulation

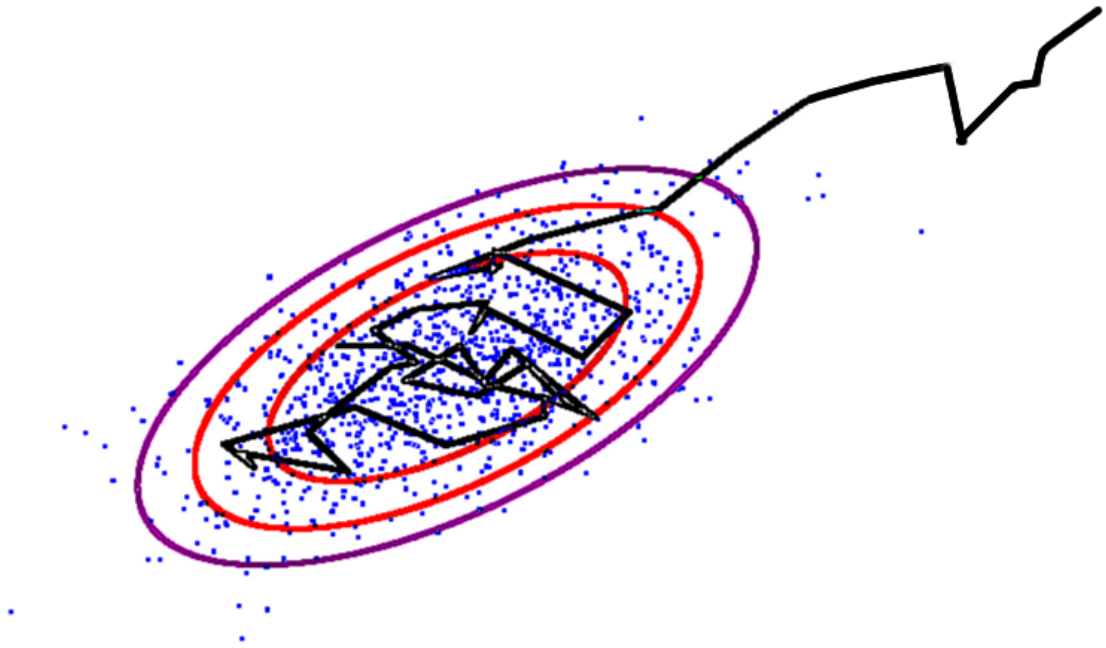
Instead we "Monte Carlo" (MC) methods to generate random deviates in the right ratios from the target posterior called "draws" or samples. We then use these draws to approximate our distribution and make inference statements (estimates, CIs, etc).

We can also use the draws to calculate the posterior distribution of **any function of our estimated parameters**. As the number of draws/samples gets large we can approximate these quantities arbitrarily high precision.

Markov Chain MC (MCMC)

The most commonly used numerical algorithm for generating posterior samples is MCMC.

A Markov Chain is a randomly generated sequence of numbers where each draw depends on the one immediately preceding it $\rightarrow$ random walk.



Plot – Ian Murray (http://mlg.eng.cam.ac.uk/zoubin/tut06/mcmc.pdf)

Gibbs Sampling

One specific algorithm that is commonly used is Gibbs Sampling.

Gibbs sampling leverages the *conditional* distributions of parameters to generate samples by proposing them one at a time. This is the algorithm implemented in the popular Bayesian packages BUGS, WinBUGS, and JAGS/`rjags`.

We will treat Gibbs sampling and other of the numerical methods as mostly "black boxes". We'll learn to diagnose output from these later on in the practical component.

What do we do with Posterior Samples?

We can treat the draws much like we would data:

- Calculate posterior summaries (mean, median, mode, etc) just like we would a data sample
- Calculate precision of the summaries (e.g., sample variance)
- CIs via quantiles (order statistics of the data) or HPD intervals (using `CODA` package in R)

If the samples are parameters in a complex model, we can plug them all in, one at a time to get a range of possible predictions from the model (we'll see this in the practical bit, later on).

How do we compare models?

The simplest way that we will use to compare models is via the Deviance Information Criterion (DIC). Like AIC and BIC, DIC seeks to judge a model on how well it fits, penalized by the complexity of the model.

$$DIC = D(\bar{\theta}) + 2p_D$$

where:

- Deviance: $D(\theta) = -2\log(\mathcal{L}(\theta; y)) + C$
- Penalty: $p_D = \bar{D} - D(\bar{\theta})$
- $D(\bar{\theta})$: deviance at the posterior mean of θ
- $\bar{D}$: average deviance across the posterior samples.

$\rightarrow$ Already implemented in JAGS!