Bayes Seminar: Part I

John Myles White

April 2, 2012

Statistical theory, minimally, tries to answer two questions:

- What can we learn from experience?
- How should we best learn from experience?

Bayesian statistical theory answers one additional question:

- How should we represent the knowledge we gain from experience?

- Since Hume, we've known that experience is fallible
- So how do we cope with experiences that are not perfectly informative?

A motivating example:

- Imagine that you've arrived in a new city for the first time
- It's raining on the day when you arrive
- Do you conclude that it rains every day in this new city?

- Most people will not reach this conclusion
- Some statistical methods will reach this conclusion
- We'd like to decide what is the *right* conclusion to draw
- How can we formalize the situation for mathematical analysis?

- We assume that each day it either rains or it does not

- We assume that each day it either rains or it does not
- We encode rain as a binary variable that takes on values of 0 or 1

- We assume that each day it either rains or it does not
- We encode rain as a binary variable that takes on values of 0 or 1
- We assume that this binary variable is a random variable with a probability distribution over $\{0, 1\}$

- We assume that each day it either rains or it does not
- We encode rain as a binary variable that takes on values of 0 or 1
- We assume that this binary variable is a random variable with a probability distribution over $\{0, 1\}$
- We assume that each separate day is an independent instantiation of this random variable

- Those assumptions are the core elements of most probabilistic models
- Having made them, we can provide a formal analysis

- Each individual day's weather is being modeled as a Bernoulli variable
- n days taken together are a binomial variable
- Either model has exactly one unknown parameter: p, the probability of raining
- We wish to estimate this parameter

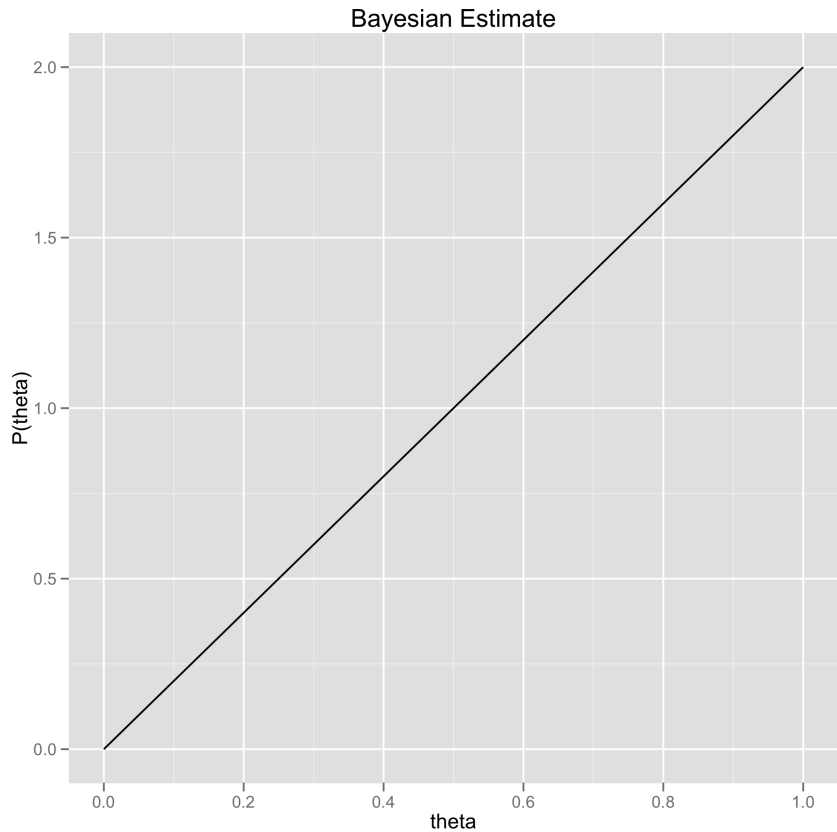
- In this context, we've answered the question of what we can learn
- p is all that we can learn
- We therefore only have to ask how to learn about p and how to represent what we learn

Some representations of p:

- Point estimation: our knowledge of p is a single value $\hat{p}$
- Interval estimation: our knowledge of p is a range, $[\hat{p}_l, \hat{p}_u]$
- Bayesian estimation: our knowledge of p is a probability distribution over $\hat{p}$'s

Some estimates given one day's worth of rain:

- Point Estimate: $\hat{p} = 1$ (MLE)
- Interval Estimate: $[\hat{p}_l, \hat{p}_u] = [0.025, 1.000]$ (95% CI)
- Bayesian Estimate:



- Throughout this seminar I'm going to focus on Bayesian estimation
- I'll contrast it with point and interval estimation
- I won't discuss hypothesis testing at all

Generalizing our example:

- We have a data set, D, of n data points: $x_1, x_2, \ldots, x_n$
- We have a probabilistic model that generates data sets
- We write the probability of a data set as $p(D) = p(x_1, x_2, \ldots, x_n)$
- In this seminar, every model will be defined by a finite number of parameters
- Instead of the single parameter, p, we'll have a list of parameters, θ
- We then write $p(x_1, x_2, \ldots, x_n|\theta)$

- If our data set is fixed, we can treat $p(x_1, x_2, \ldots, x_n|\theta)$ as a function of θ
- We'll sometimes write $L(\theta; x_1, x_2, \ldots, x_n)$ to describe this function
- This function is called the likelihood function
- L tells us the probability of seeing any specific data set if the parameters of the model were set to θ

Simplifying the likelihood function:

- Because we want to deal with arbitrarily large data sets easily using models we already have on hand, we'll typically simplify the models of data we use
- Specifically, we'll take a probabilistic model of individual data points and build up a model of data sets
- We do this by assuming that all data points in the data set have an identical probability distribution and that they are all independent of each other
- We call this the IID assumption

The IID Assumption:

- For all i, $p(x_i)$ is a single, unvarying probability distribution
- All i data points are independent samples from this constant underlying distribution
- With these assumptions, any data set has the property that

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2)\ldots p(x_n) = \prod_{i=1}^{n} p(x_i)$$

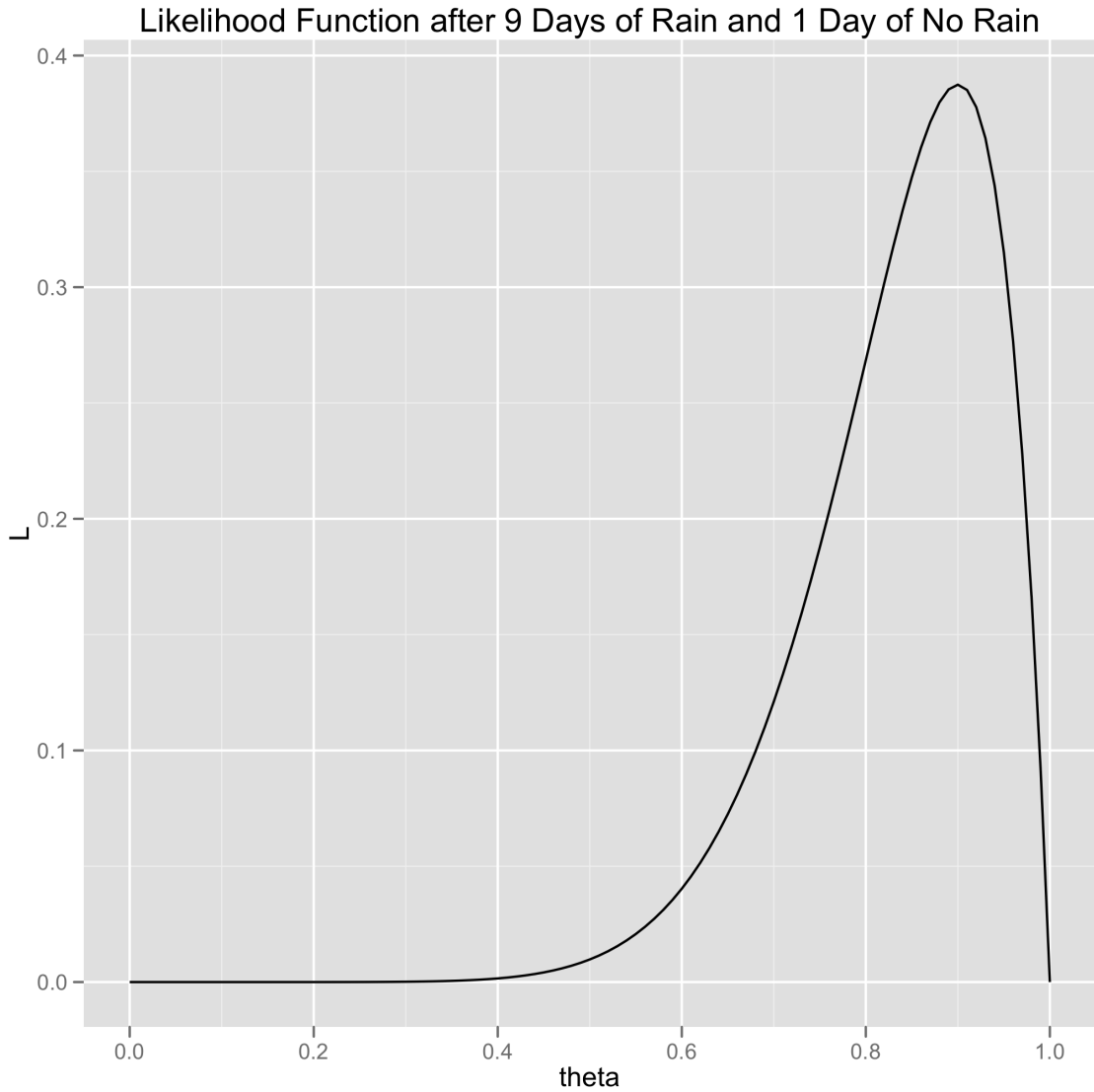
- This factorization makes certain types of mathematical analysis very simple

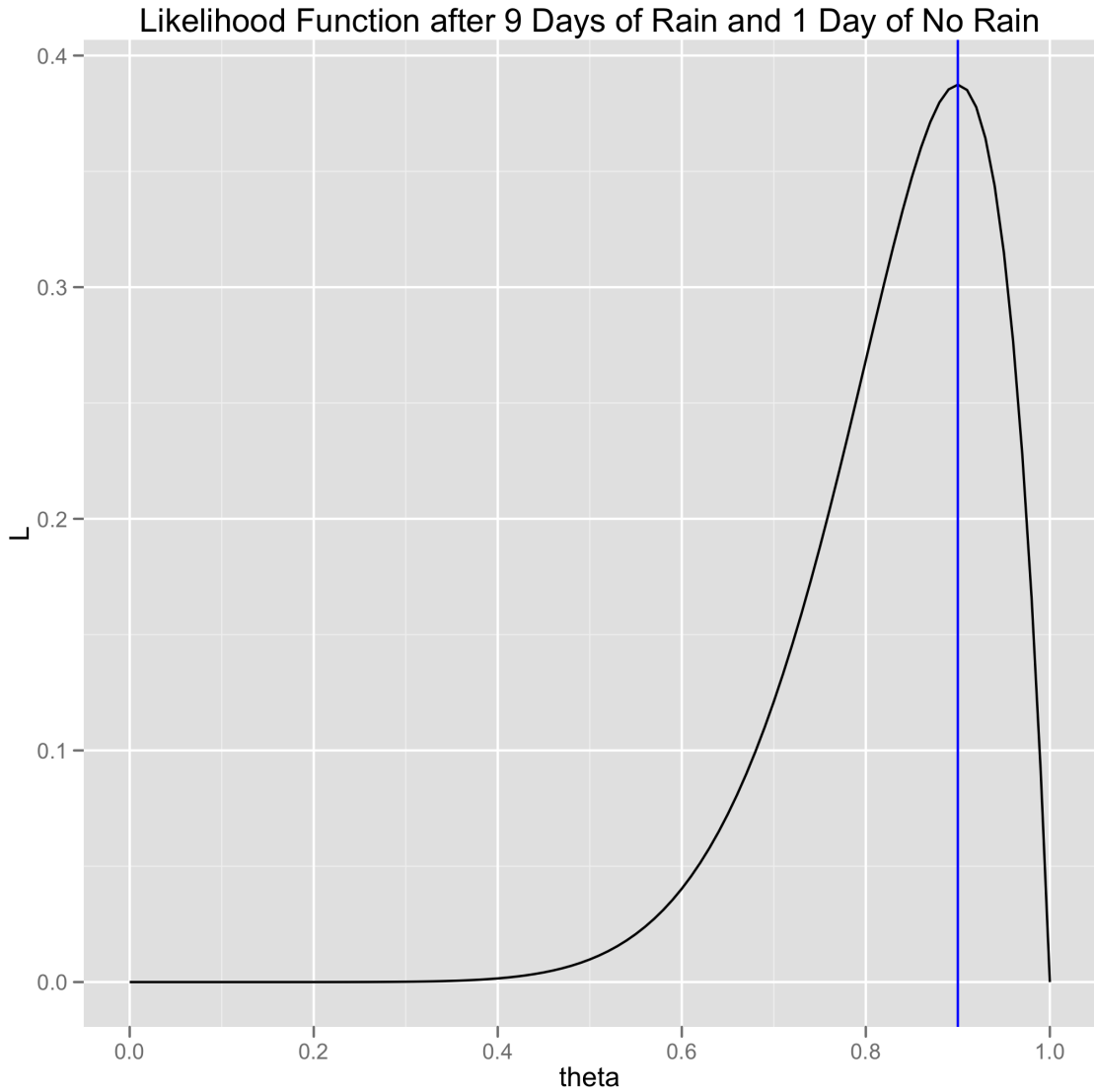
- Let's see how these assumptions play out for data about many days worth of weather
- We'll assume we've seen n days of weather
- We'll assume it rained on $n - 1$ days and did not rain on only 1 day

- The probability of rain on any given day is θ
- The occurrence of rain on each of the n days is independent
- The probability of our data set given θ is thus

$$\binom{n}{n-1}\theta^{n-1}(1-\theta)$$

- Given our data, how should we estimate θ?
- We'll start by graphing the likelihood function as we change θ
- For this example, let's assume that $n = 10$


Likelihood Function after 9 Days of Rain and 1 Day of No Rain
L
0.0
0.1
0.2
0.3
0.4
0.0
0.2
0.4
0.6
0.8
1.0
theta


Likelihood Function after 9 Days of Rain and 1 Day of No Rain
0.4
0.3
0.2
0.1
0.0
L
0.0
0.2
0.4
0.6
0.8
1.0
theta

- We've seen rain on 90% of days
- The likelihood function has a single peak at $\theta = 0.9$
- We might guess that we can estimate θ by maximizing the likelihood function

- Estimating parameters by maximizing the likelihood function is a good general strategy
- It also seems intuitively reasonable:

  If we had to predict the future from our model, we should use parameter values that would increase our chances of predicting the data from the past

- For our example model, maximizing the likelihood function can be done analytically
- In other models, computational methods are required

- Note that the maximum of

$$\binom{n-1}{n}\theta^{n-1}(1-\theta)$$

- occurs at the same place as the maximum of

$$\theta^{n-1}(1-\theta)$$

- Then note that the maximum of

$$\theta^{n-1}(1-\theta)$$

- occurs at the same place as the maximum of

$$\log[\theta^{n-1}(1-\theta)] = (n-1)\log[\theta] + \log[1-\theta]$$

- The default parameter estimation strategy of maximizing the log likelihood function comes from Fisher
- Fisher demonstrated that it was a very powerful general strategy for estimating parameters

- Prior to Fisher's work, statistics often involved the ad hoc construction of methods for estimating parameters
- Let's review those ideas, because they're important for thinking critically about Bayesian estimation

- We'll call any function of our data a statistic
- Any statistic designed to give a point estimate of θ is an estimator

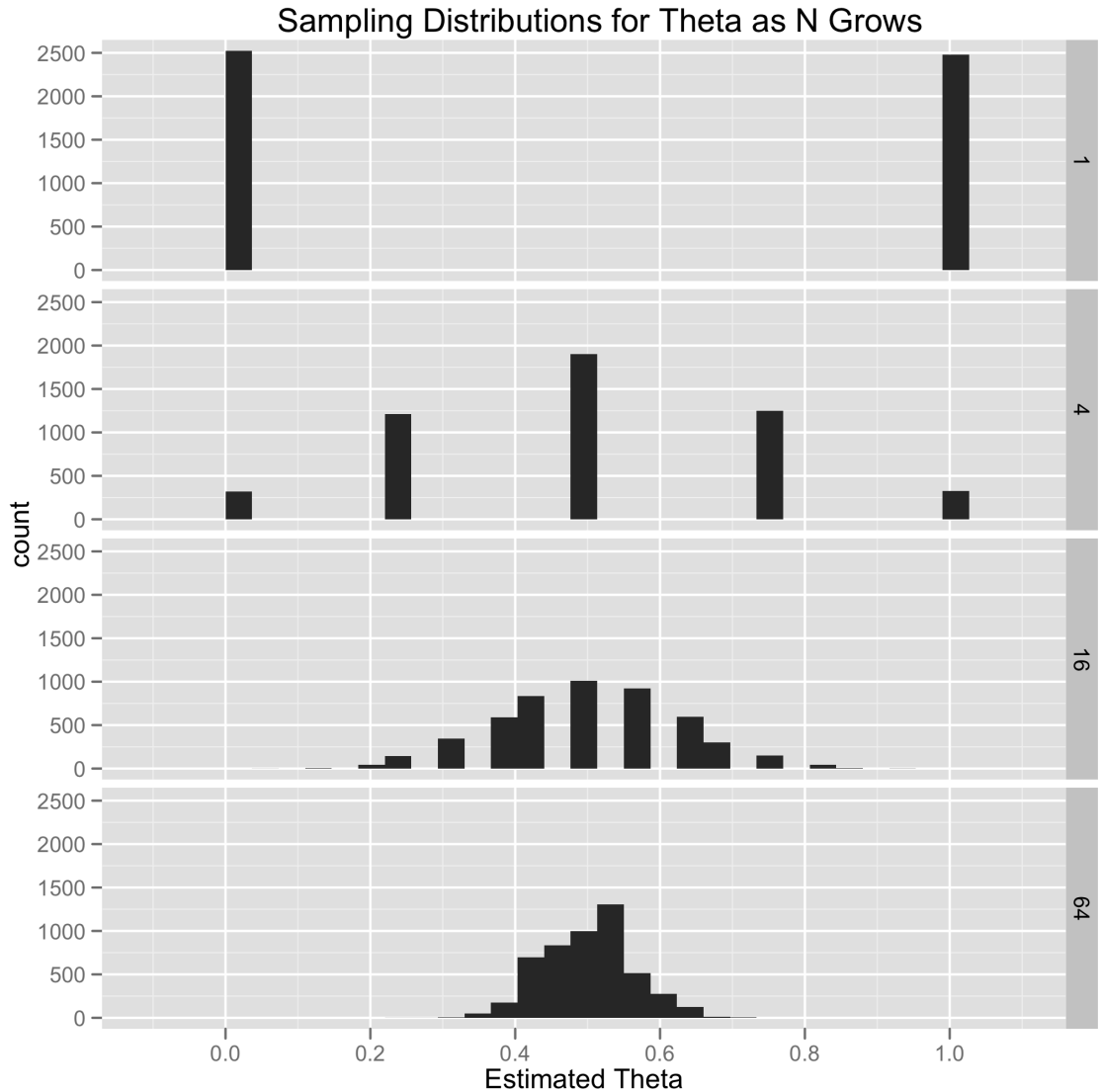
- Suppose we have data that comes from a normal distribution with mean μ
- How can we estimate μ?

- One estimator is the mean of the data
- Another is the median

- How can we tell which of these two estimators works better?

Sampling distribution analysis:

- Suppose that θ is fixed
- We repeatedly sample random data sets from our model
- We compute our estimator on the resulting data sets
- We look at the distribution of our estimator after repeated sampling


Sampling Distributions for Theta as N Grows
count
2500
2000
1500
1000
500
0
1
4
16
64
0.0
0.2
0.4
0.6
0.8
1.0
Estimated Theta

- We then analyze the quality of an estimator by analyzing its sampling distribution

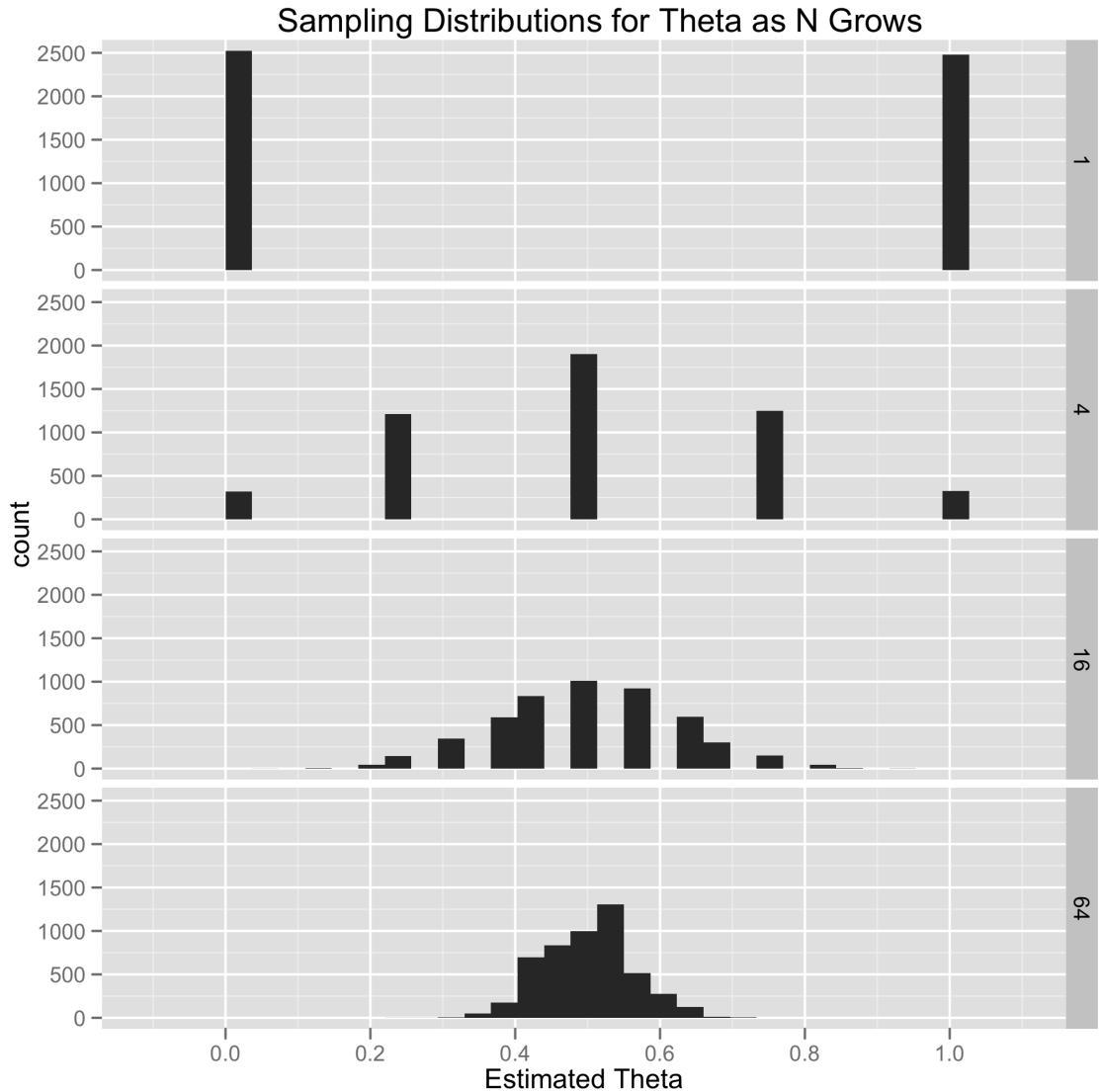
Three criteria for selecting estimators are particularly popular:

- Bias
- Variance
- Consistency

- The bias of an estimator, $\hat{\theta}$, is

$$\mathbb{E}[\hat{\theta} - \theta] = \mathbb{E}[\hat{\theta}] - \theta$$

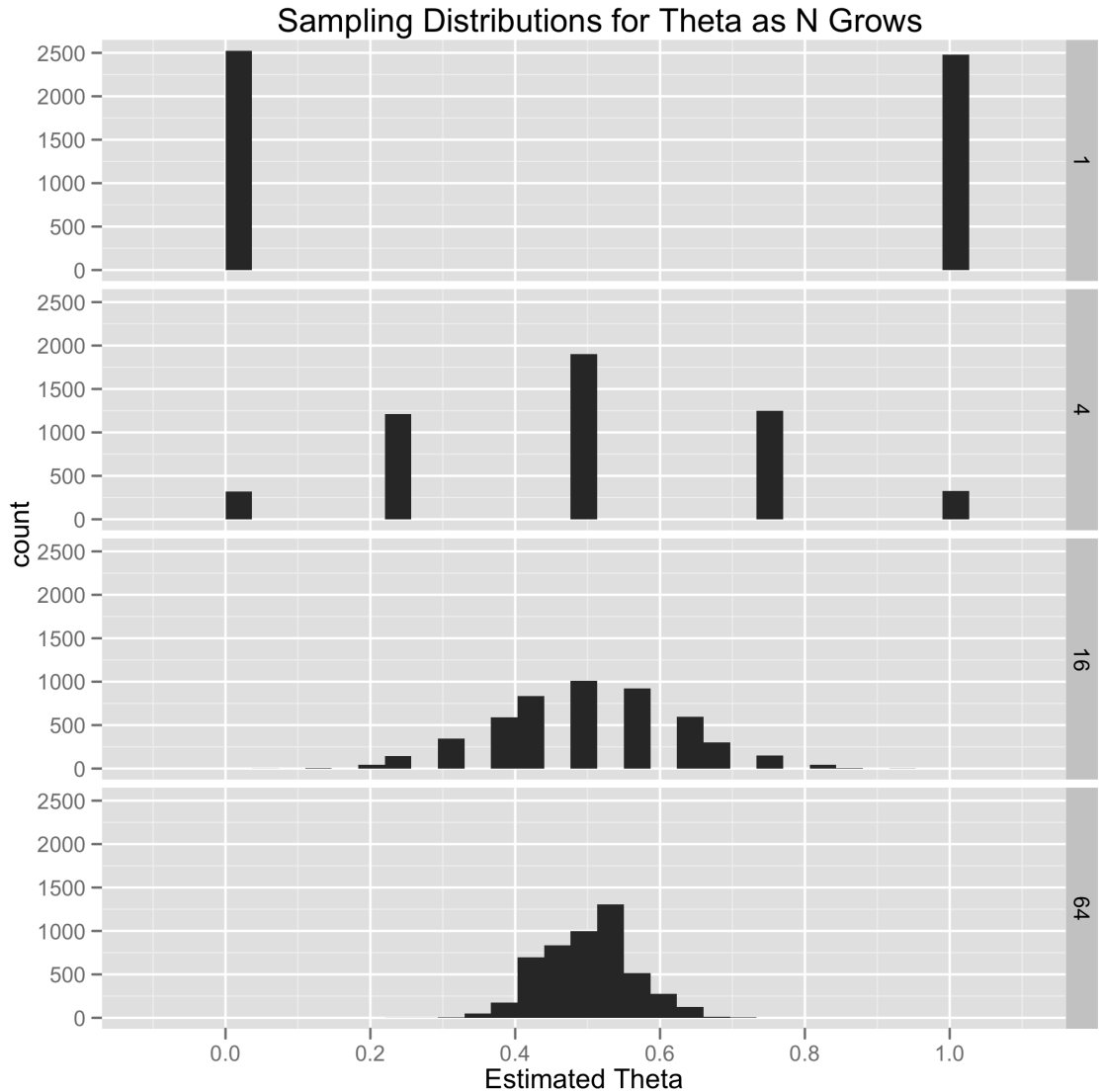
- An unbiased estimator is one for which the mean of the sampling distribution is θ
- In short, an unbiased estimator's expected value is θ


Sampling Distributions for Theta as N Grows
count
2500
2000
1500
1000
500
0
1
4
16
64
0.0
0.2
0.4
0.6
0.8
1.0
Estimated Theta

- The variance of an estimator, $\hat{\theta}$, is

$$\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

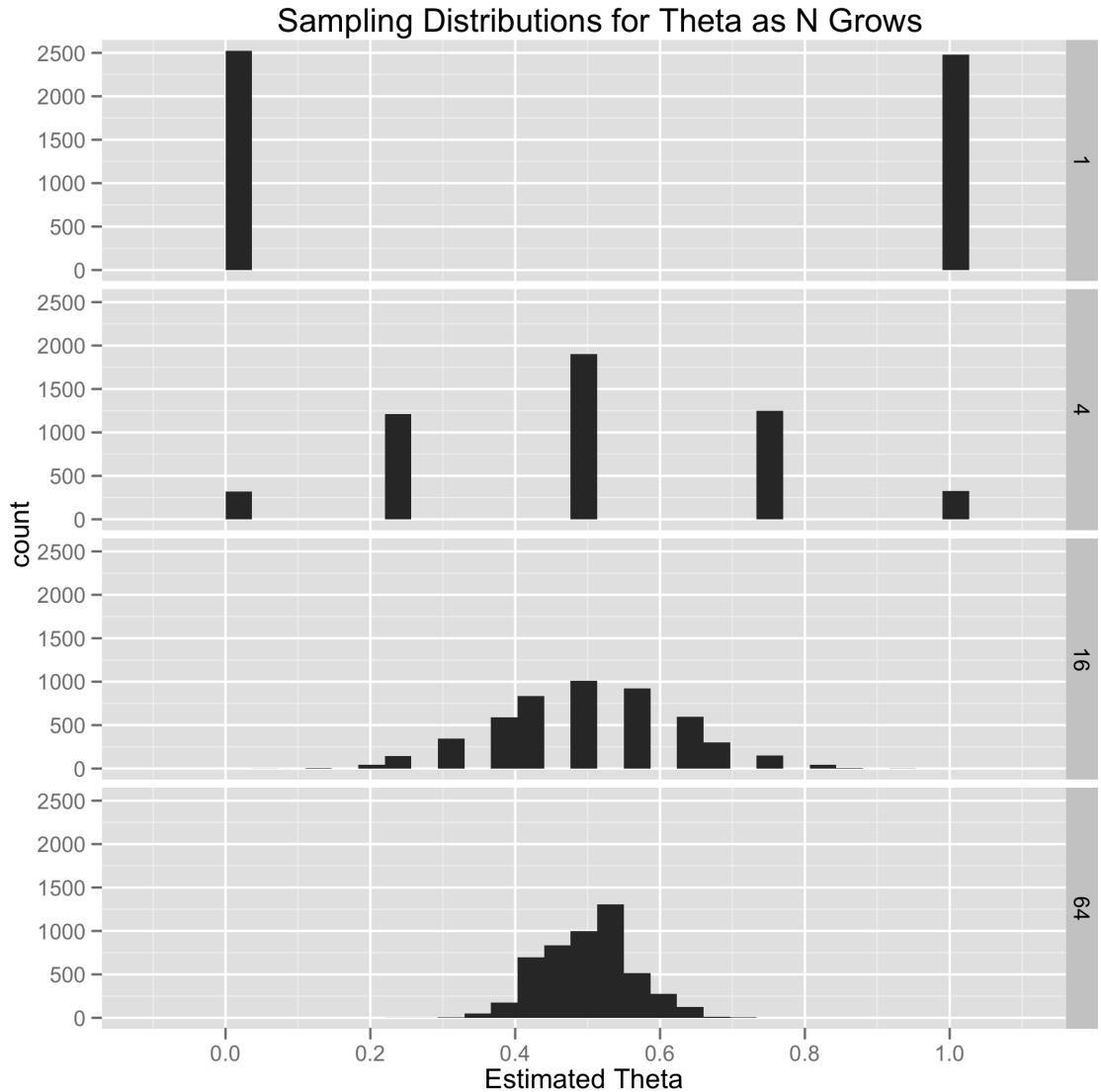
- If the estimator is unbiased, the variance is the expected Euclidean distance between $\hat{\theta}$ and θ
- In short, a low variance estimator is one that is typically close to $\mathbb{E}[\theta]$


Sampling Distributions for Theta as N Grows
count
2500
2000
1500
1000
500
0
1
4
16
64
0.0
0.2
0.4
0.6
0.8
1.0
Estimated Theta

- An estimator, $\hat{\theta}$, is consistent if

$$\lim_{n \to \infty} \hat{\theta} = \theta$$

- An consistent estimator gets closer to θ as we get more data
- To be consistent, the bias and variance must both go to 0 as n grows


Sampling Distributions for Theta as N Grows
count
2500
2000
1500
1000
500
0
1
4
16
64
0.0
0.2
0.4
0.6
0.8
1.0
Estimated Theta

- Fisher made the MLE popular by showing that typically:
  - The MLE is unbiased
  - The MLE is consistent
  - The MLE is asymptotically the lowest variance estimator

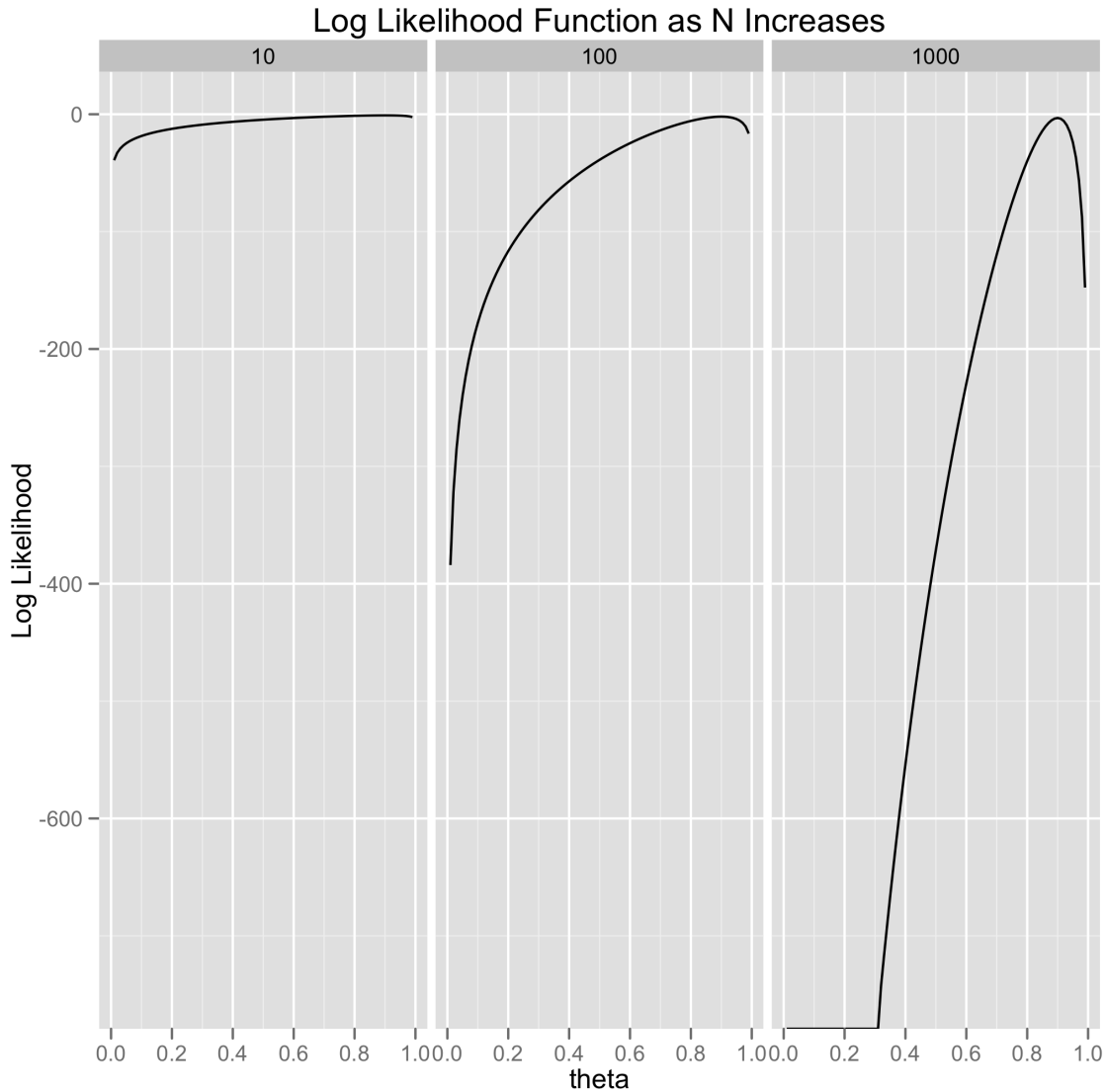
- From the 1920's until the 1950's, the MLE was king
- In the 50's, James-Stein published a simple problem for which the MLE was a bad estimator

- Since the 50's, interest has grown in alternative estimation strategies
- It is now clear that unbiased estimators are sometimes bad in practice
- It is now clear that finite sample behavior is not the same as asymptotic behavior

- Enter the Bayesian estimation strategy
- In many practical problems, Bayesian methods are biased
- But, in many practical problems, Bayesian methods have lower variance than MLE methods
- Moreover, bias and variance can often be explicitly traded off

- Let's go back to the likelihood function
- We've seen one way of using it to construct estimators
- What other information can we get from it?

- We can get a measure of our uncertainty about θ
- We do this by looking at the curvature of the likelihood function
- Formally, this is the Fisher information


Log Likelihood Function as N Increases
10
100
1000
0
-200
-400
-600
Log Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
theta

- We've now seen that we can learn a lot from the likelihood function
  - The peak gives us an estimator
  - The spread gives us a sense of uncertainty

- Pushing on these ideas, we might wish to have a single numeric language for representing our knowledge of θ that organizes the information in the likelihood function
- For a Bayesian that language is probability theory
- We'd like to say that we think that $\hat{\theta}$ is the most probable value for θ
- We'd also like to say that there is a 95% chance that θ's value is in some interval, $[a, b]$

- To motivate that language, let's start with a very loose treatment of the Cox axioms
- We'll follow Jaynes' treatment

The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

- James Clerk Maxwell

Traditional logic:

- If A then B
- A
- B

- How do we extend this approach to situations in which A is not certain?

The Cox Axioms a la Jaynes:

- This approach is a heuristic for thinking about how reasoning *should* work in a hypothetical robot
- It may not be fully rigorous
- It is unquestionably *not* a description of human reasoning *does* works

Axiom I: Degrees of Plausibility are represented by real numbers

Axiom II: Qualitative correspondence with common sense

- If $(A|C') > (A|C)$ and $(B|AC') = (B|AC)$
- Then $(AB|C') \geq (AB|C)$ and $(\bar{A}|C') < (\bar{A}|C)$

Axiom III: Consistency

- 3a: If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result
- 3b: The robot always takes into account all of the evidence it has relevant to the question. It does not arbitrarily ignore some of the information, basing its conclusions only on what remains. In other words, the robot is completely non-ideological
- 3c: The robot always represents equivalent states of knowledge by equivalent probability assignments. That is, if in two problems the robot's state of knowledge is the same (except perhaps for the labelling of the propositions), then it must assign the same plausibilities in both

- We can satisfy these demands if we represent our belief in statements about θ using probability theory
- So let's represent our beliefs using probability distributions

- We start with a probability distribution over θ: $p(\theta)$
- We call this distribution our prior
- It is a prior because it represents our beliefs before we see data

- We wish to calculate our belief distribution after seeing data, $D = x_1, \ldots, x_n$: $p(\theta | x_1, \ldots, x_n)$
- We call this distribution our posterior, $p(\theta | D)$
- It is a posterior because it represents our beliefs after we see data

- We can calculate our posterior using Bayes' theorem:

$$p(\theta|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{p(x_1, \ldots, x_n)}$$

- $p(x_1, \ldots, x_n)$ is called the evidence. It is a constant wrt θ
- Therefore

$$p(\theta|x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n|\theta)p(\theta)$$

In words:

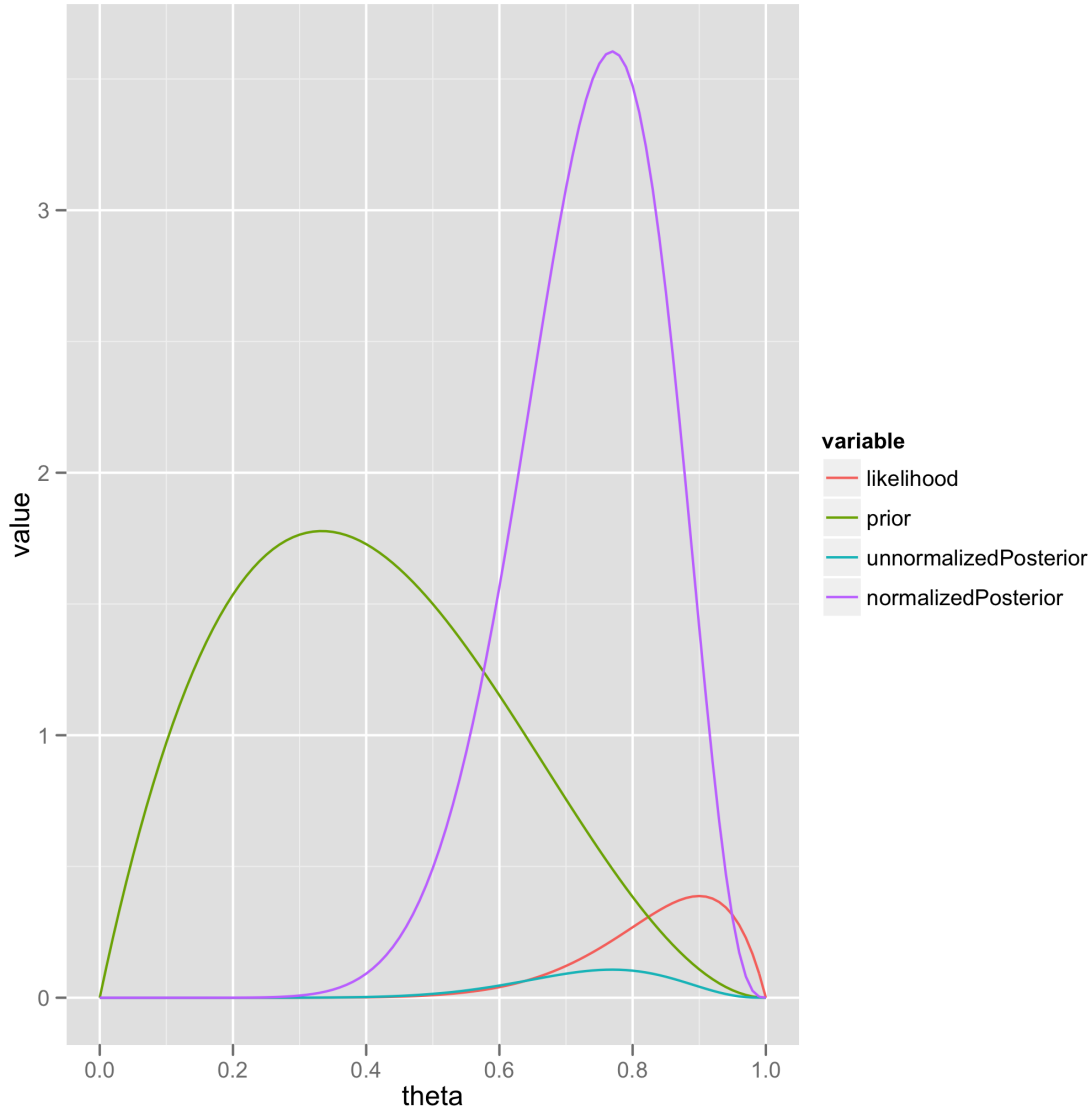
- Up to a scaling factor, the value of our posterior at a point θ^* is the product of the likelihood function evaluated at θ^* and the prior evaluated at θ^*
- If our prior is flat, the posterior's shape is the likelihood function's shape

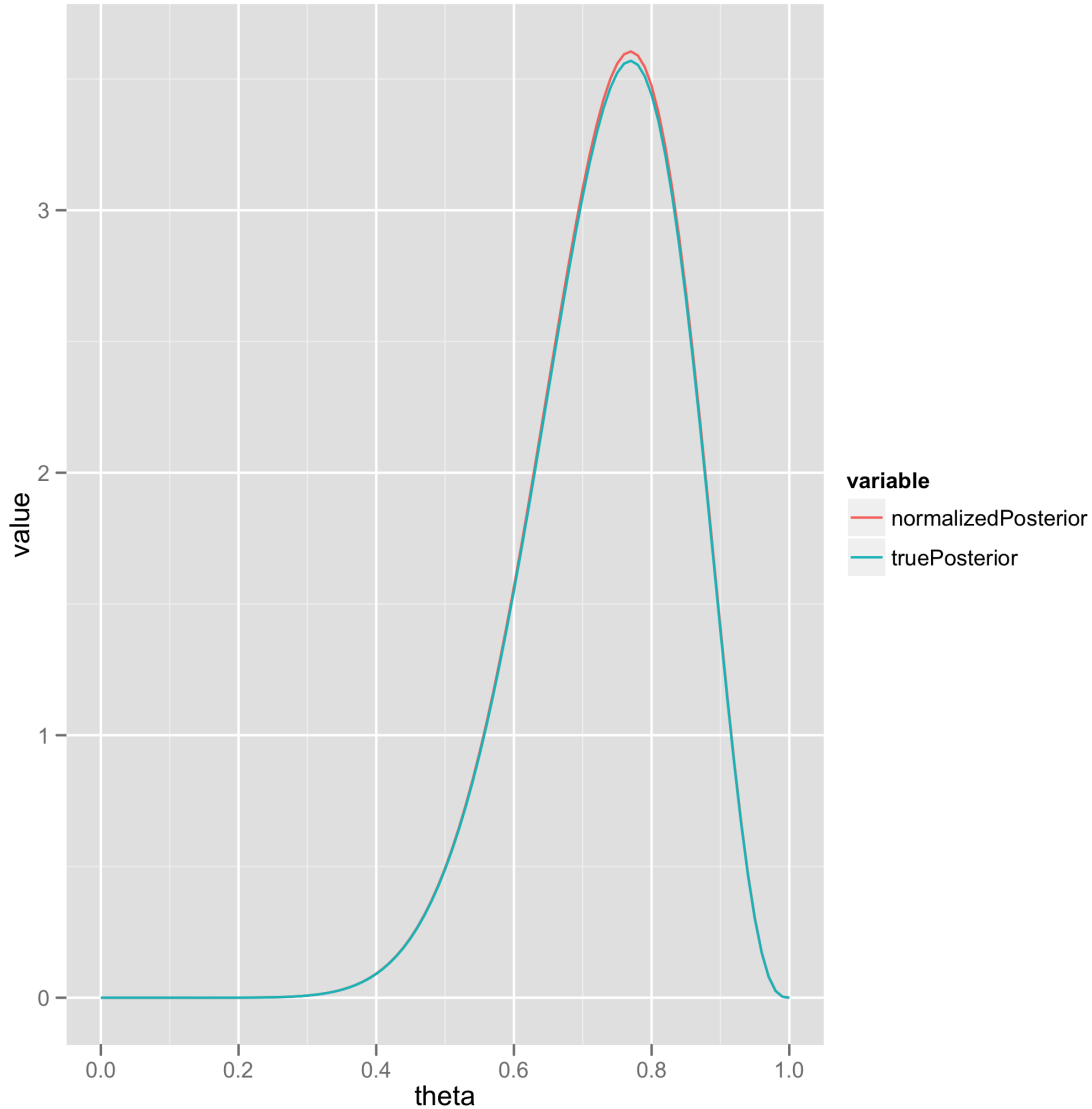
- In practice calculating the posterior is hard because the evidence can be impossible to calculate analytically
- But we can usually make good approximations
- And, in some special circumstances, we can get analytic solutions

Approximation strategy I:

- We want the posterior only at n points on a grid
- We find the unnormalized posterior by multiplying the likelihood and prior
- We evaluate the approximate evidence by summing the unnormalized posterior
- We divide the unnormalized posterior by the approximate evidence to get a proper probability distribution

- Suppose we start with a prior biased towards low values of θ
- Then we see data for which the likelihood function supports high values of θ
- The posterior describes our reconciliation of these two pieces of information


value
theta
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
variable
likelihood
prior
unnormalizedPosterior
normalizedPosterior


value
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
theta
variable
normalizedPosterior
truePosterior

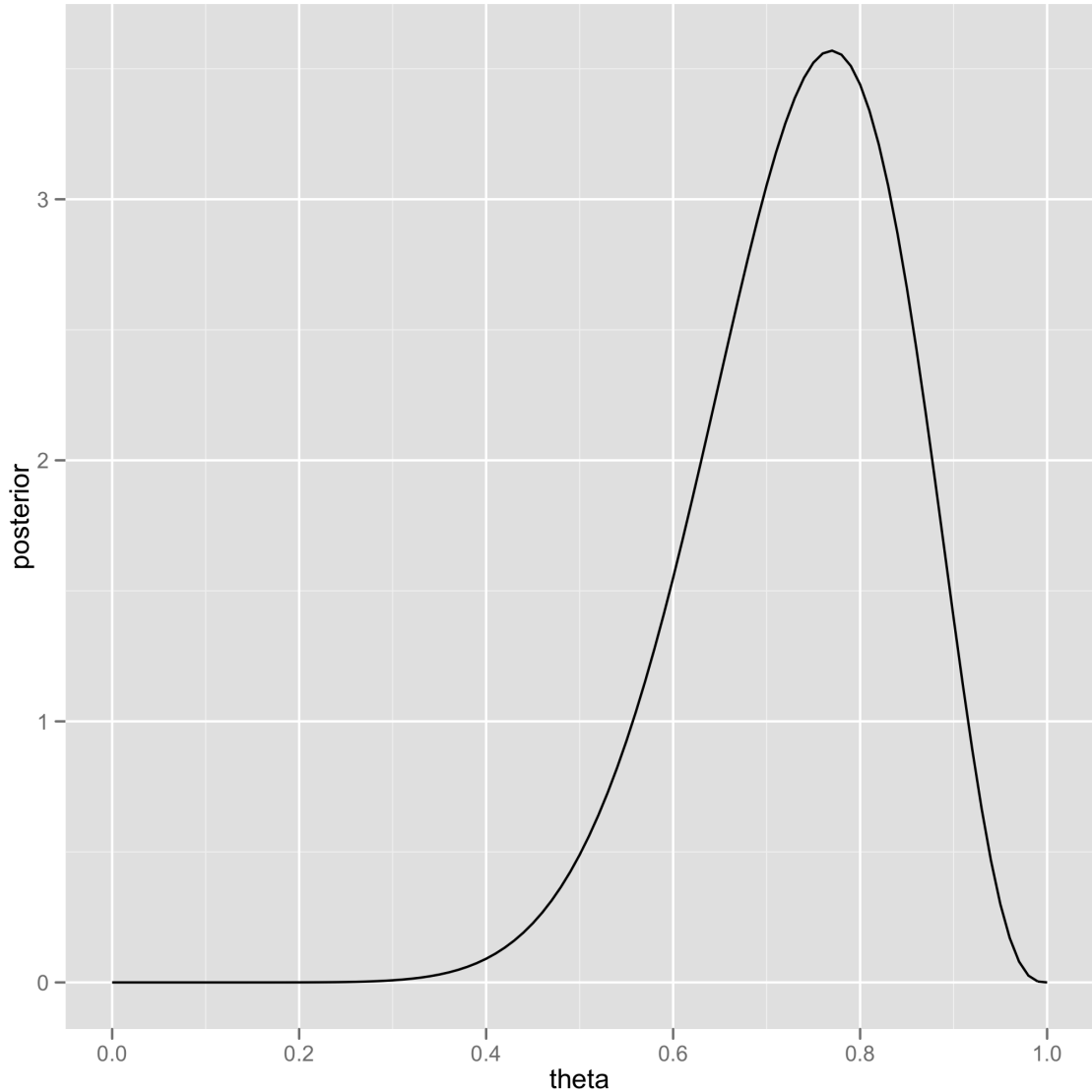
- Where does grid approximation go wrong?
- Computational time and space is exponential in number of parameters in θ
- Unclear how fine resolution of grid must be

- There are important cases where analytic techniques work
- Given an appropriate prior, the posterior can have a closed form solution
- If both the prior and posterior have a constant functional form F, the prior is called conjugate

- In our binomial example, a prior called the beta prior is conjugate to the binomial likelihood function
- The beta prior, $B(\alpha, \beta)$ has two parameters, α and β
- α is the number of 1's previously seen
- β is the number of 0's previously seen
- After x more 1's and y more 0's, the posterior has the form

$$B(\alpha + x, \beta + y)$$

- Let's work further with our example
- Our prior will be a $B(2, 3)$ distribution
- We'll suppose we've seen 9 more days of rain and 1 new dry day
- Our posterior is then a $B(11, 4)$ distribution


posterior
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
theta

- Visualizing the full posterior can tell us a lot
- What other things can we do?

- We can extract various point estimates from the posterior
  - Means
  - Medians
  - Modes
- We can compute interval estimates
  - 95% probability intervals using quantiles

- When we use the mode of the posterior as our estimate we call it the MAP
- MAP stands for Maximum A Posteriori
- It is the Bayesian analogue to the MLE

Using the MAP via maximizing its log value is like penalized maximum likelihood:

$$\log[p(\theta|D)] = \log[L(\theta|D)p(\theta)] = \log[L(\theta|D)] + \log[p(\theta)]$$

- Instead of finding the MAP, we can extract other values:
  - The mean of the posterior
  - The median of the posterior
  - Quantiles of the posterior
- To decide which value to use, we can use decision theory

- In practice, from now on, we're just going to use the mean and median of the posterior as point estimates

- In modern Bayesian statistics, we use conjugacy whenever we can
- Otherwise, we use MCMC techniques
- For me, that means I always use MCMC

- MCMC techniques draw samples from the posterior
- There are therefore Monte Carlo methods
- As such, their quality increases when we take more samples
- To find each sample, a Markov chain is used
- Therefore MCMC = Markov chain Monte Carlo methods

- For many problems, MCMC analysis can be completely automatic
- Part II of the seminar will work through many examples of applied Bayesian computation
- We'll use an MCMC tool called BUGS that fully automates MCMC sampling
- If you want to learn how MCMC works under the hood, there are many good books