Introduction to Bayesian Statistics

Environmental Data Analytics

Dr. Alix I. Gitelman

Statistics Department
gitelman@science.oregonstate.edu

July 2015

Overview

1. An introduction: philosophy, basic manipulations, priors, interpretation
2. Computational methods for estimation: MCMC
3. Models, examples, practice
4. More on modeling, wrap-up

What's a Paradigm?!

"a philosophical and theoretical framework of a scientific school or discipline within which theories, laws, and generalizations and the experiments performed in support of them are formulated"

–Merriam-Webster online

"the set of common beliefs and agreements shared between scientists about how problems should be understood and addressed"

–Thomas Kuhn

What's the Frequentist Paradigm?

Specifically, what are the "beliefs and agreements" about how statistical problems should be "understood and addressed?"

Let's back up even further: what are "statistical problems?"

- **estimation**
- **hypothesis testing**

How do frequentists do estimation and testing (in general)?

- **using sampling distributions**

What's the Frequentist Paradigm?

Let's go deeper still: what is a sampling distribution?

...somewhere along the way the word "frequency" ought to come up, no?

- Remember that a sampling distribution of a particular statistic, T_n, is the relative frequency distribution of T_n constructed from repeated samples of size n from the population of interest.

So, the frequentist paradigm relies on the notion of relative frequency probability for dealing with statistical problems.

Relative Frequency Probability

Suppose that A is one possible outcome among a finite set of possible outcomes of some experiment E. Then $Pr(A)$ is defined as the relative frequency of A's occurrence in an infinite sequence of repeated experiments, E.

There's a lot of hypothetical stuff in that definition:

1. We need to assume that it's possible to run E more than once, let alone an infinite number of times!
2. We need to wait around for an infinity of trials!

How Frequentist Modeling Works

In the frequentist approach we:

1. Sample data from an unknown population distribution
2. Condition on a true, fixed, *unknown* quantity or set of quantities
3. Determine how unusual a sample we have relative to all other possible samples

What's the Bayesian Paradigm?

Like the frequentist paradigm, the Bayesian paradigm has at its heart a notion of probability.

- Subjective probability is defined in terms of bets: A probability p attached to an event E is defined as the fraction $p \in [0, 1]$ at which you would bet $\$p$ for a return of \$1 if E occurs.

Notice that this means that I might assign a different probability to a certain event than you might.

What's the Bayesian Paradigm?

In the Bayesian approach, we:

1. Specify a prior model for the parameter(s) of interest
2. Obtain data from a relevant population
3. Condition on the observed data to update our prior model to a **posterior model** that we use to make inference.

Notice the similarity to the Scientific Method

Frequentist vs Bayesian

In terms of making inferences there are two essential differences between classical (frequentist) and Bayesian statistics:

1. How we think about parameters (fixed versus random)
2. How we think about probability (long-run frequency versus "subjective")

To a frequentist, parameters are "true, fixed, unknown" quantities.

By contrast, a Bayesian models uncertainty about parameters. In that regard, parameters are thought of as random.

Frequentist vs Bayesian

To Frequentists:

- Parameters are fixed
- Data are random

To Bayesians:

- Parameters are random
- Data are obtained through some random mechanism, but in an analysis, they are treated as fixed

Lichen Presence/Absence

Surveying the diversity and abundance of lichens in forests can be useful for several reasons:

1. To assist in classification of stands as "old-growth"
2. To evaluate climate conditions and effects
3. To evaluate stand health vis-a-vis airborne pollution.

In an oversimplification of a lichen survey, we'll look at the presence/absence of one species, *lobaria oregana* or "lettuce lichen," which is relatively common in the Oregon Cascades.

Lettuce Lichen

Lichen Presence/Absence

Suppose in a given forest stand in the Oregon Cascades, we obtain a sample of $n = 57$ spatially distinct trees.

1. Each tree is evaluated for the presence/absence of a particular lichen species, *lobaria oregana*, or "lettuce lichen."
2. We'll assume that being "spatially distinct" is enough to ensure that the observations are statistically independent.
3. In all, $X = 22$ of the trees have the lichen present.

The Binomial Probability Distribution

A reasonable probability model for these data is the Binomial probability mass function:

$$Pr(X = x) = \binom{57}{x} \pi^x (1 - \pi)^{57-x}$$

for $x = 0, 1, \ldots, 57$ and $0 \leq \pi \leq 1$.

Here, π is the population probability of presence.

In general, the Binomial probability distribution is written,

$$P(X = x|\pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

for $x = 0, 1, \ldots, n$ and $0 \leq \pi \leq 1$.

Frequentist Approach

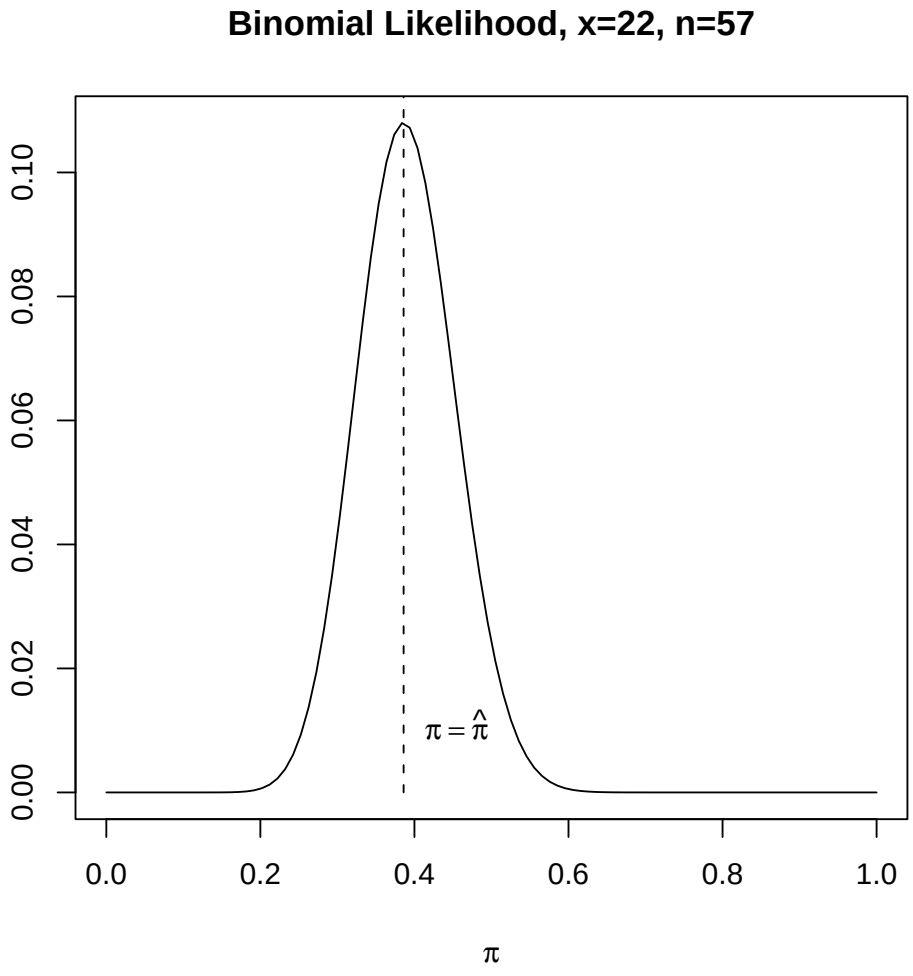
How would a frequentist analyze these data?

- estimate π with $\hat{\pi} = 22/57 = 0.3860$
- find $SE(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})/57} = 0.0645$
- identify the sampling distribution for $\hat{\pi}$.

Some questions:

1. Why is $\hat{\pi} = x/n$ a good estimate for π?
2. What do we use as a sampling distribution?

The Maximum Likelihood Estimate



Frequentist Approach

R R code to create the plot:

```
pi.vec = seq(0,1,length=100)
bin.fun = function(p) choose(57,22)*p^22*(1-p)^(57-22)

plot(pi.vec,bin.fun(pi.vec),type="l",xlab=expression(pi),ylab="")
lines(c(22/57,22/57),c(0,.13),lty=2)
text(locator(1),expression(pi == hat(pi)))
title("Binomial Likelihood, x=22, n=57")
```

What do we use as the sampling distribution for $\hat{\pi}$, and why?!

```
data:  22 out of 57, null probability 0.5
X-squared = 2.5263, df = 1, p-value = 0.1120
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.2629116 0.5243842
```

Bayesian Approach

Here, we'll again use the Binomial distribution function for X:

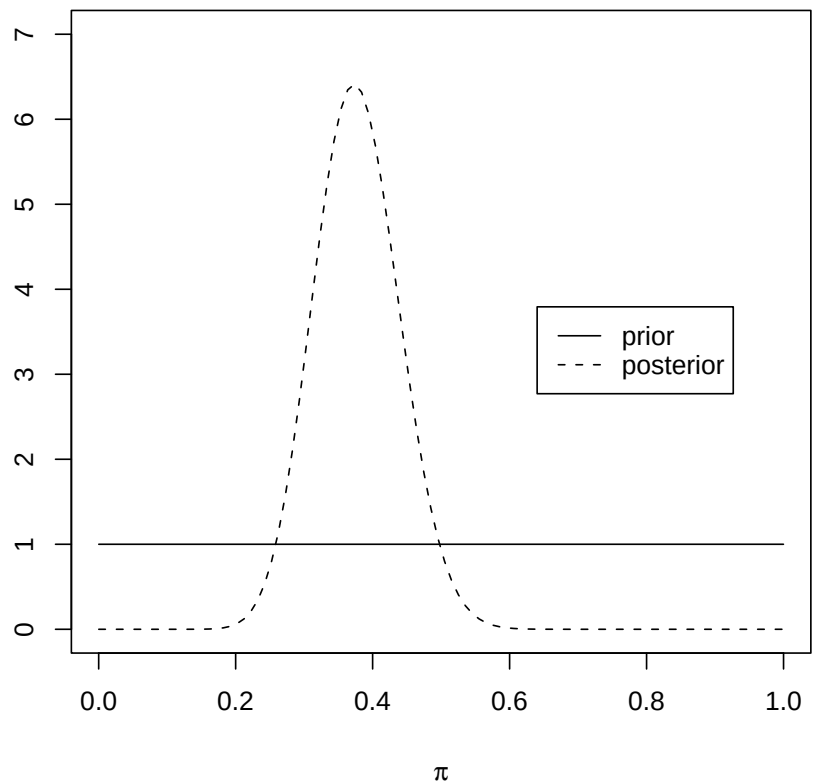
$$P(X = x|\pi) = \binom{n}{x}\pi^x(1-\pi)^{57-x}$$

And now, we have to specify a prior probability model for π.

How to do this? Can we draw a picture?

Bayesian Approach

Using Bayes Theorem (details later), I now combine the likelihood model and uniform (or non-informative or reference) prior model to obtain the **posterior distribution** for π; namely, $f(\pi|X)$.



Bayesian Approach

Using this **posterior distribution**, I can report the **posterior mean** of π and a **95% posterior interval** for π (this is the Bayesian analog to the frequentist confidence interval):

- The posterior mean, $\text{mean}(\pi|X) = \tilde{\pi} = 0.40$.
- A 95% posterior interval is (0.28,0.52).

Compare this to the frequentist estimate, $\hat{\pi} = 0.38$ and confidence interval: (0.26,0.52).

Interpretations

For the frequentist confidence interval:

> *"In 95% of repeated samples, the 95% confidence interval for π will cover π."*

Notice a couple of things:

1. We don't say anything about **this particular** interval, we just have to make a general statement about hypothetical intervals like this one.

2. The language "will cover π" reflects the fact that the confidence interval is a probability statement about the sample proportion, $\hat{\pi}$; it's not a probability statement about π.

Interpretations

For the Bayesian posterior interval:

> *"Under a uniform prior for π, and given the observed data, the probability that π is between 0.28 and 0.52 is 95%."*

Some things to notice:

1. This is a statement about **this** interval, not some hypothetical collection of intervals.

2. The probability statement is about π, the parameter we actually want to make inference about!

A Comment on Notation

Let's suppose that $X_1, \ldots, X_n \overset{ind}{\sim} f(x; \theta)$.

Then the joint probability distribution of $(X_1, \ldots, X_n) \equiv \mathbf{X}$ is just

$$f(\mathbf{X}; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

This joint pdf is also called the likelihood function, and in frequentist statistics, we often want to maximize the likelihood in θ, and therefore we write the likelihood as a function of θ:

$$L(\theta|\mathbf{X}) = \prod_{i=1}^{n} f(x_i; \theta)$$

In Bayesian statistics, we really do think of the likelihood function as the joint probability distribution of the data, given θ, and so we typically write the likelihood as $L(\mathbf{X}|\theta)$ or $L(\mathbf{X}; \theta)$.

Bayes Theorem: Discrete Probabilities

Suppose that A and B are discrete events where each can take a finite number of possible values; that is, $A = a$ for $a \in \{a_1, a_2, \ldots, a_k\}$ and $B = b$ for $b \in \{b_1, b_2, \ldots, b_\ell\}$.

Bayes Theorem:

$$\begin{aligned} P(A = a|B = b) &= \frac{P(B = b|A = a)P(A = a)}{\sum_{j=1}^{k} P(B = b|A = a_j)P(A = a_j)} \\ &= \frac{P(B = b|A = a)P(A = a)}{P(B = b)} \end{aligned}$$

provided that $P(B = b) \neq 0$.

Bayes Theorem: Discrete Probabilities

Bayes Theorem:

$$P(A = a|B = b) = \frac{P(B = b|A = a)P(A = a)}{P(B = b)} \tag{1}$$

provided that $P(B = b) \neq 0$.

Notice that the **order of conditioning** changes from the right hand side $[P(B = b|A = a)]$ to the left hand side $[P(A = a|B = b)]$

Bayes Theorem: An Example

You've likely seen Bayes theorem in the context of a medical screening test: Suppose are given a screening test for a particular disease, the result is positive, and now you what to know the probability that you actually have the disease.

- Suppose $Pr(disease) = Pr(D) = 0.0012$. This is the incidence rate.
- Further suppose $Pr(positive|disease) = Pr(P|D) = 0.95$. This is the true positive rate (also called the sensitivity of the screening test)
- A suppose $Pr(positive|disease^c) = Pr(P|D^c) = 0.001$. This is the false positive rate. For good screening tests, this should be very low.

Bayes Theorem: An Example

Applying Bayes theorem:

$$\begin{aligned} Pr(D|P) &= \frac{P(P|D)P(D)}{P(P|D)P(D) + P(P|D^c)P(D^c)} \\ &= \frac{0.95(0.0012)}{0.95(0.0012) + 0.001(1 - 0.0012)} \\ &= 0.53. \end{aligned}$$

The combination of a rare disease and a non-zero false positive rate make this probability smaller than you might think at first.

The bottom line: before hitting the panic button after a positive screening test, be sure you understand the incidence rate of the disease or condition as well as the true and false positive rates.

Bayes Theorem: Generalization

Bayes rule also applies to probability distribution functions.

Some notation:

- $L(\mathbf{X}|\theta)$ denotes the likelihood function of a sample of data, $\mathbf{X}$, conditional on some parameter(s), θ.
- Remember that a likelihood function is the joint pdf (or pmf) of a sample of data.
- If $X_1, \ldots, X_n \overset{iid}{\sim} f(x_i|\theta)$, then:

$$L(\mathbf{X}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

is the likelihood function.

Bayes Theorem: Generalization

Bayes Theorem applied to probability distribution (mass) functions:

General Version of Bayes Theorem:

$$f(\theta|\mathbf{X}) = \frac{L(\mathbf{X}|\theta)f(\theta)}{\int L(\mathbf{X}|\theta)f(\theta)d\theta}$$

provided some regularity conditions are met.

The regularity conditions are to ensure that the denominator is finite (this is essentially analogous to the condition that $P(B = b) \neq 0$ in the discrete case).

Components of Bayes Theorem

To repeat, Bayes Theorem:

$$f(\theta|\mathbf{X}) = \frac{L(\mathbf{X}|\theta)f(\theta)}{\int L(\mathbf{X}|\theta)f(\theta)d\theta}$$

provided some regularity conditions are met.

Here, the distribution function, $f(\theta|\mathbf{X})$, is called the **posterior distribution** of θ given $\mathbf{X}$.

And $f(\theta)$ is called the **prior distribution** of θ.

Bayesian Modeling

In a Bayesian model:

1. We specify a probability model for the observed data—this is the likelihood function;
2. we specify a prior probability model for the parameters of interest (i.e., the parameters of the likelihood model);
3. and, we use Bayes theorem to combine the likelihood and the prior to obtain the posterior distribution.

The posterior distribution is the primary inferential tool in Bayesian statistics—with it, we can report posterior probabilities about the parameters of interest.

Back To Bayes Theorem

Bayes Theorem:

$$f(\theta|\mathbf{X}) = \frac{L(\mathbf{X}|\theta)f(\theta)}{\int L(\mathbf{X}|\theta)f(\theta)d\theta}$$

provided some regularity conditions are met.

We have to deal with the denominator term, which can be daunting, especially if θ is multi-dimensional.

Notice, however, that $\int L(\mathbf{X}|\theta)f(\theta)d\theta$ does not depend on θ—that's the whole point of the integration, to get rid of θ.

Normalizing Constant

Therefore, in Bayes Theorem, since the left hand side, $f(\theta|\mathbf{X})$, is a function of θ, and the denominator on the right hand side is a constant with respect to θ, we can think of the denominator as a **scaling factor** or **normalizing constant** and write:

$$\begin{aligned} f(\theta|\mathbf{X}) &= \frac{L(\mathbf{X}|\theta)f(\theta)}{g(\mathbf{X})} \\ &\propto L(\mathbf{X}|\theta)f(\theta) \end{aligned}$$

That is: **posterior** $\propto$ **likelihood** $\times$ **prior**.

The symbol $\propto$ is read: "is proportional to"

Back to the Binomial Problem

For our Binomial problem with $X = 22$ and $n = 57$ we have

$$L(X = 22|\pi) = \binom{57}{22}\pi^{22}(1-\pi)^{35}$$

And, using a Uniform(0,1) prior for π, we have

$$f(\pi) = 1 \text{ for } 0 < \pi < 1.$$

So that

$$\begin{aligned} f(\pi|X = 22) &= \frac{\binom{57}{22}\pi^{22}(1-\pi)^{35}}{\int_0^1 \binom{57}{22}\pi^{22}(1-\pi)^{35}d\pi} \\ &\propto \binom{57}{22}\pi^{22}(1-\pi)^{35} \end{aligned}$$

The Beta pdf

If a random variable, θ, follows a Beta(α, β) distribution, then the pdf of θ is:

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

for $\theta \in [0,1]$ and $\alpha, \beta > 0$.

The gamma function, $\Gamma(\alpha)$ is defined for $\alpha > 0$:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$$

- $\Gamma(\alpha+1) = \alpha\Gamma(\alpha), \alpha > 0$
- $\Gamma(n) = (n-1)!, n \in \mathbb{Z}^+$
- $\Gamma(1/2) = \sqrt{\pi}, \pi = 3.14....$

The Beta pdf

The **kernel** of a density function is just that part of the function that remains when constants are discarded.

The Beta pdf is a function of the parameter, θ, and so α and β are constants.

So in

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

The **kernel** is:

$$\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The Binomial Problem

Our result from a few slides back:

$$f(\pi|X = 22) \quad \propto \quad \underbrace{\pi^{22}(1-\pi)^{35}}_{\text{kernel of a Beta distribution}}$$

What are the parameters of this Beta posterior distribution?

- solve $\alpha_n - 1 = 22$ for α_n, so that $\alpha_n = 23$
- solve $\beta_n - 1 = 35$ for β_n, so that $\beta_n = 36$

The Beta Posterior

What we just figured out is that if $X \sim Bin(n, \theta)$, and we use a uniform (non-informative) prior distribution for θ, then:

$$(\theta | X = x) \sim \text{Beta}(\alpha_n, \beta_n)$$

were $\alpha_n = x + 1$ and $\beta_n = n - x + 1$.

Going back to the lettuce lichen example, the posterior distribution of the probability of lichen presence is Beta(23,36), which is precisely the plot I showed earlier.

Properties of a Beta Random Variable

If $\theta \sim \text{Beta}(\alpha, \beta)$, then:

1. $E[\theta] = \frac{\alpha}{\alpha+\beta}$
2. $var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
3. And, we can use R to obtain percentiles of any Beta pdf.

The Posterior Mean

If $\theta|X \sim \text{Beta}(x+1, n-x+1)$, then the posterior mean of θ given $X = x$ is:

$$\frac{x+1}{(n-x+1)+(x+1)} = \frac{x+1}{n+2}$$

Notice that this falls in between $\hat{\theta} = x/n$ and $1/2$.

We'll see this again and again:

- A posterior mean is often a compromise between the prior mean (in this case $1/2$) and the data mean (in this case x/n).

Making Predictions

Prediction of future values or events is a big part of Statistics, and it is straightforward in the Bayesian paradigm.

Suppose $Y_1, \ldots, Y_n \overset{ind}{\sim} f(y|\theta)$, and that $\tilde{Y}$ is a new value to be predicted that is independent of the observed data.

The **posterior predictive distribution** of $\tilde{Y}$ is

$$\begin{aligned} f(\tilde{y}|\mathbf{y}) &= \int f(\tilde{y}, \theta|\mathbf{y}) d\theta \\ &= \int f(\tilde{y}|\theta, \mathbf{y}) f(\theta|\mathbf{y}) d\theta \\ &= \int f(\tilde{y}|\theta) f(\theta|\mathbf{y}) d\theta \end{aligned}$$

Informative Beta Priors

Suppose we have data, $X \sim \text{Bin}(n, \theta)$.

The Beta family of priors is **conjugate** for the binomial likelihood, so we'll take

$$\theta \sim \text{Beta}(\alpha_0, \beta_0).$$

- Remember that the prior mean of θ is therefore:

$$E[\theta] = \frac{\alpha_0}{\alpha_0 + \beta_0}$$

- So, α_0 is like a prior number of successes and $(\alpha_0 + \beta_0)$ is like a prior sample size.
- We can use these facts to formulate a prior for θ.

Informative Beta Priors

Suppose that you think a reasonable prior guess at a value for θ is 0.25, but that you're only willing to "bet" 4 prior observations on it.

Using this information, you can solve

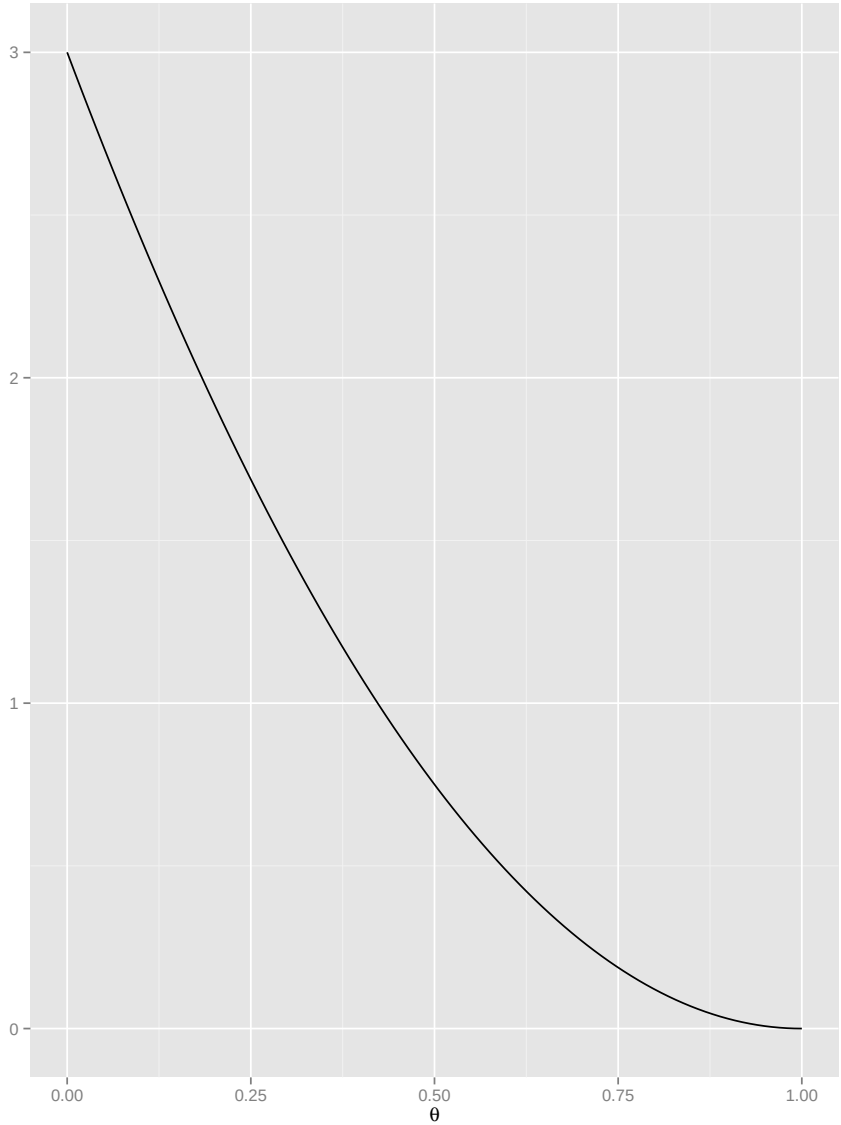
$$\begin{aligned} \frac{\alpha_0}{\alpha_0 + \beta_0} &= 0.25 \\ \alpha_0 + \beta_0 &= 4 \end{aligned}$$

to get $\alpha_0 = 1$ and $\beta_0 = 3$.

You can look at this prior distribution using:

```
th = seq(0,1,length=100)
density = dbeta(th,1,3)
df1 <- data.frame(th,density)
p  <- ggplot(df1,aes(theta,density))
p + geom_line() + xlab(expression(theta)) + ylab("")
```

The Beta(1,3) Prior



What Comes Next?

For many one-parameter problems like the binomial problem we just discussed, we can obtain analytical (or exact) solutions for the posterior distribution.

This is even true for some multi-parameter problems (e.g., Normal data with unknown mean and variance)

But, for many interesting problems, there aren't analytical solutions, and so we must rely on numerical and computational methods for estimation.

More this afternoon.