Introduction to Bayesian Statistics

Rebecca C. Steorts

Agenda

- Motivations
- Traditional inference
- Bayesian inference
- Bernoulli, Beta, and Binomial distributions
- Posterior of Beta-Binomial
- Example with 2012 election data
- Marginal likelihood
- Posterior Prediction

Motivations for Bayesian statistics

- Understanding social networks
- Predicting elections
- Estimating the size of a population
- Estimating hard to reach populations (domains)

Traditional inference

You are given **data** X and there is an **unknown parameter** you wish to estimate θ

How would you estimate θ?

- Find an unbiased estimator of θ.
- Find the maximum likelihood estimate (MLE) of θ by looking at the likelihood of the data.
- If you cannot remember the definition of an unbiased estimator or the MLE, review these before our next class.

Bayesian inference

Bayesian methods trace its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace discovered what we now call **Bayes' Theorem**

- $p(x \mid \theta)$ likelihood
- $p(\theta)$ prior
- $p(\theta \mid x)$ posterior
- $p(x)$ marginal distribution

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

Bernoulli distribution

The Bernoulli distribution is very common due to binary outcomes.

- Consider flipping a coin (heads or tails).
- We can represent this a binary random variable where the probability of heads is θ and the probability of tails is $1 - \theta$.

The write the random variable as $X \sim \text{Bernoulli}(\theta)\mathbb{1}(0 < \theta < 1)$

It follows that the likelihood is

$$p(x \mid \theta) = \theta^x (1-\theta)^{(1-x)} \mathbb{1}(0 < \theta < 1).$$

- Exercise: what is the mean and the variance of X?

Binomial distribution

- Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Then for $x_1, \ldots, x_n \in \{0, 1\}$ what is the likelihood?

Likelihood

$$\begin{aligned} p(x_{1:n}|\theta) &= \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid \theta) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid \theta) \\ &= \prod_{i=1}^{n} p(x_i|\theta) \\ &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}. \end{aligned}$$

Beta distribution

Given $a, b > 0$, we write $\theta \sim \text{Beta}(a, b)$ to mean that θ has pdf

$$p(\theta) = \text{Beta}(\theta|a, b) = \frac{1}{B(a, b)} \theta^{a-1}(1-\theta)^{b-1} \mathbb{1}(0 < \theta < 1),$$

i.e., $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$ on the interval from 0 to 1.

- Here,
$$B(a, b) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$
.
- The mean is $E(\theta) = \int \theta \, p(\theta) d\theta = a/(a + b)$.

Notation

- $\propto$: means "proportional to"
- $x_{1:n}$ denotes $x_1, \ldots, x_n$

Posterior of Binomial-Beta

Lets derive the posterior of $\theta \mid x_{1:n}$

$$
\begin{aligned}
& p(\theta|x_{1:n}) \\
& \propto p(x_{1:n}|\theta)p(\theta) \\
& = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}I(0<\theta<1) \\
& \propto \theta^{a+\sum x_i-1}(1-\theta)^{b+n-\sum x_i-1}I(0<\theta<1) \\
& \propto Beta\big(\theta \mid a+\sum x_i,\, b+n-\sum x_i\big).
\end{aligned}
$$

Approval ratings of Obama

What is the proportion of people that approve of President Obama in PA? - We take a random sample of 10 people in PA and find that 6 approve of President Obama. - The national approval rating (Zogby poll) of President Obama in mid-September 2015 was 45%. We'll assume that in PA his approval rating is approximately 50%. - Based on this prior information, we'll use a Beta prior for θ and we'll choose a and b.

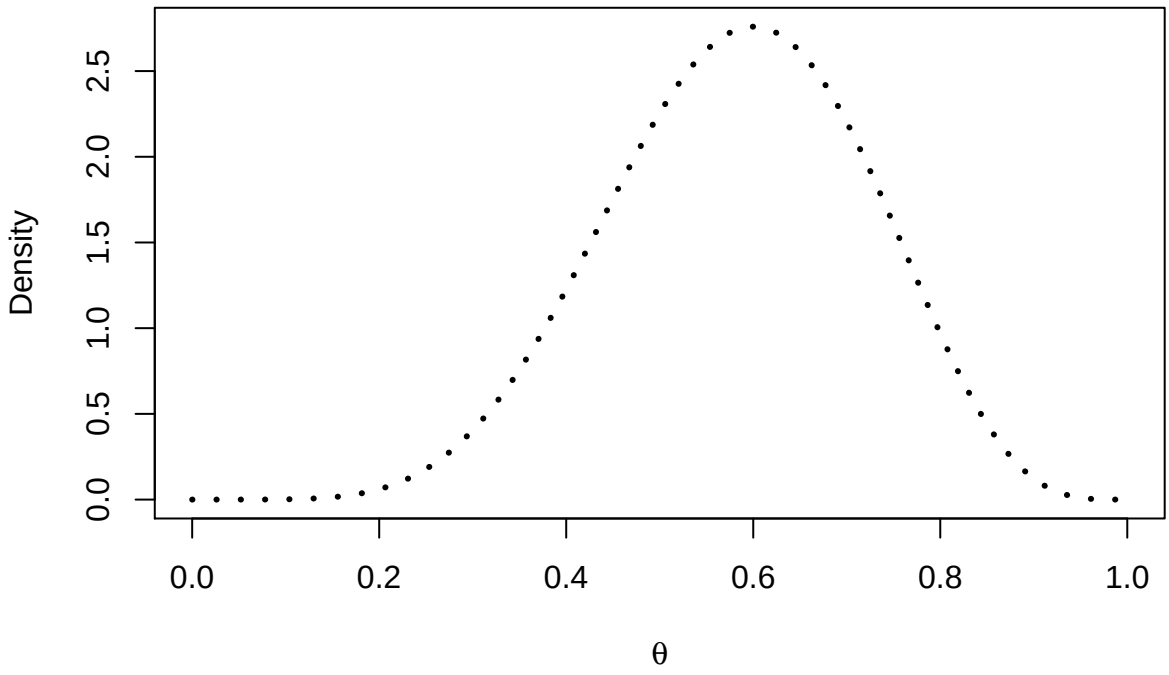
Obama Example

```
n = 10
#center theta at 1/2 and spread at 0.04
a = 21/8
b = 0.04
th = seq(0,1, length=500)
x = 6

# we set the likelihood, prior, and posteriors with THETA as
# the sequence that we plot on the x-axis.
# Beta(c,d) refers to shape parameter
like = dbeta(th, x+1, n-x+1)
prior = dbeta(th, a, b)
post = dbeta(th, x+a, n-x+b)
```

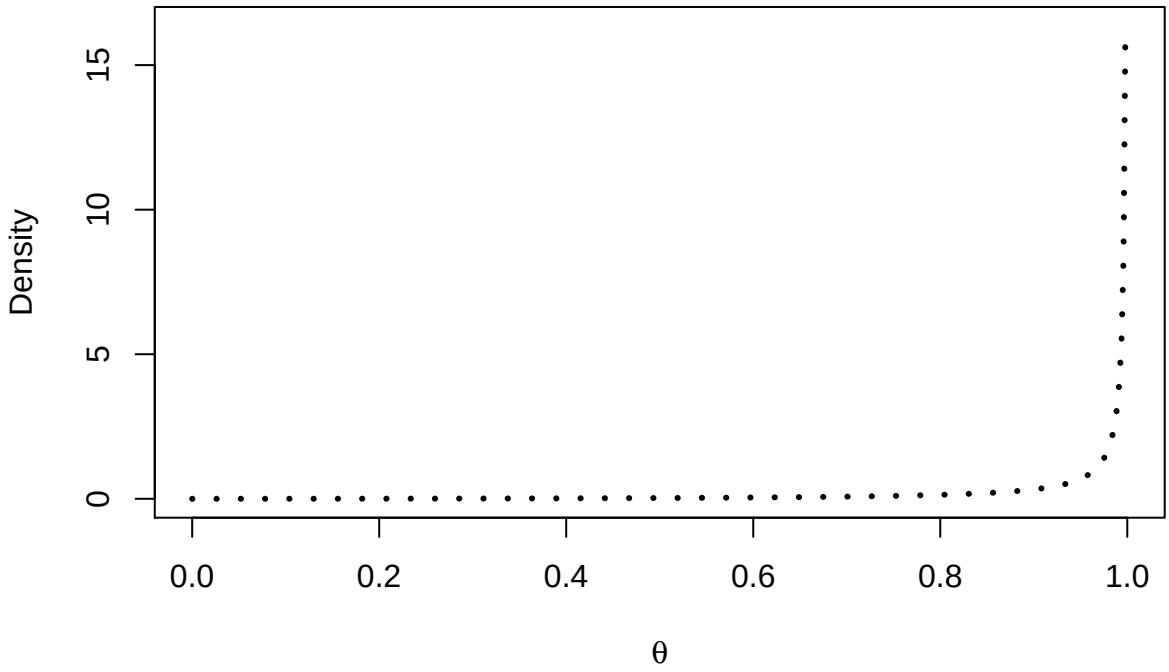
Likelihood

```
plot(th, like, type='l', ylab = "Density", lty = 3, lwd = 3, xlab = expression(theta))
```



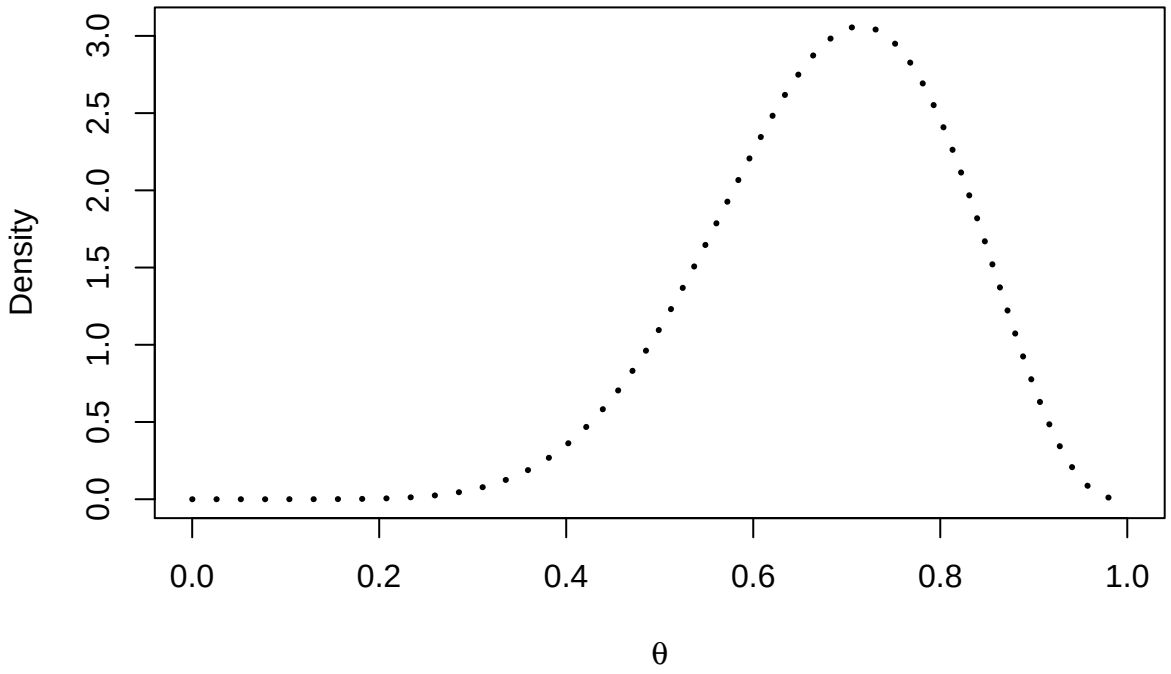
Prior

```
plot(th, prior, type='l', ylab = "Density", lty = 3, lwd = 3, xlab = expression(theta))
```



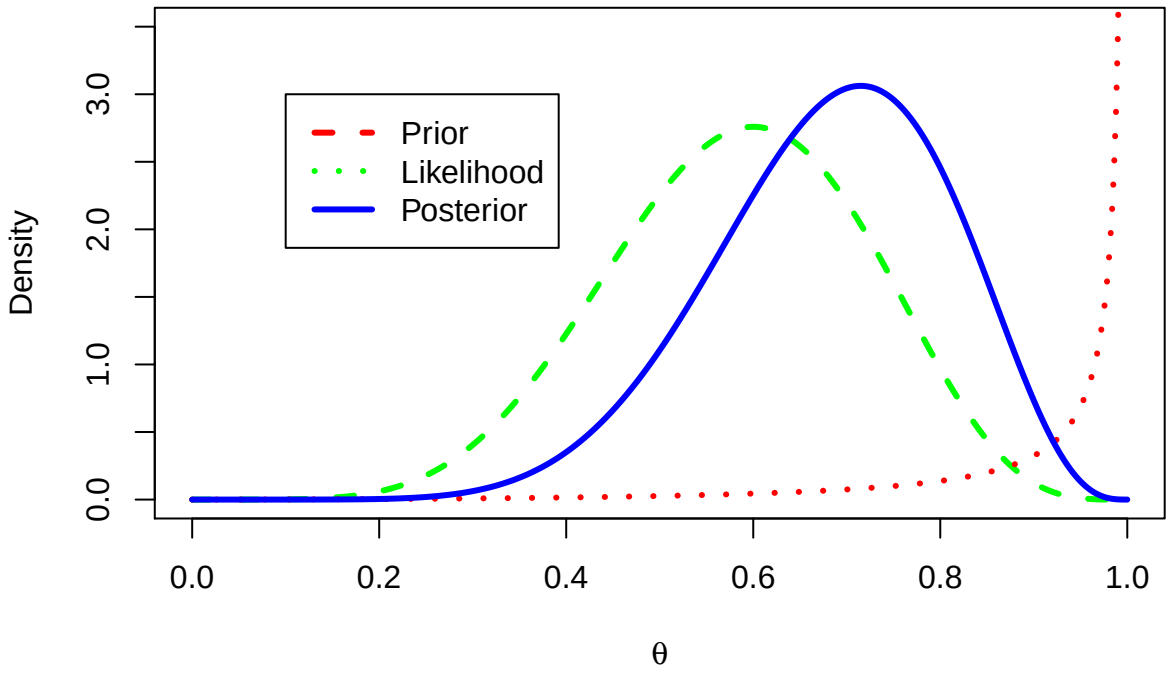
Posterior

```
plot(th, post, type='l', ylab = "Density", lty = 3, lwd = 3, xlab = expression(theta))
```



Likelihood, Prior, and Posterior

```
plot(th, like, type = "l", ylab = "Density", xlab = expression(theta), lty = 2, lwd = 3,
     col = "green",ylim = c(0,3.5) )
lines(th, prior, lty = 3, lwd = 3, col= "red")
lines(th, post, lty=1, lwd = 3, col= "blue")
legend(0.1,3, c("Prior", "Likelihood", "Posterior"), lty=c(2,3,1), lwd=c(3,3,3),
       col = c("red", "green", "blue"))
```



Cast of characters

- Observed data: x
- Note this could consist of many data points, e.g., $x = x_{1:n} = (x_1, \ldots, x_n)$.

likelihood	$p(x\|\theta)$
prior	$p(\theta)$
posterior	$p(\theta\|x)$
marginal likelihood	$p(x)$
posterior predictive	$p(x_{n+1}\|x_{1:n})$
loss function	$\ell(s, a)$
posterior expected loss	$\rho(a, x)$
risk / frequentist risk	$R(\theta, \delta)$
integrated risk	$r(\delta)$

Marginal likelihood

The **marginal likelihood** is

$$p(x) = \int p(x|\theta)p(\theta)\, d\theta$$

- What is the marginal likelihood for the Bernoulli-Beta?

Posterior predictive distribution

- We may wish to predict a new data point x_{n+1}
- We assume that $x_{1:(n+1)}$ are independent given θ

$$
\begin{aligned}
p(x_{n+1}|x_{1:n}) &= \int p(x_{n+1}, \theta|x_{1:n})\, d\theta \\
&= \int p(x_{n+1}|\theta, x_{1:n}) p(\theta|x_{1:n})\, d\theta \\
&= \int p(x_{n+1}|\theta) p(\theta|x_{1:n})\, d\theta.
\end{aligned}
$$

Example: Back to the Beta-Bernoulli

Suppose

$$\theta \sim Beta(a, b)$$

and

$$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} Bernoulli(\theta)$$

Then the marginal likelihood is

$$
\begin{aligned}
&p(x_{1:n}) \\
&= \int p(x_{1:n}|\theta) p(\theta)\, d\theta \\
&= \int_0^1 \theta^{\sum x_i} (1-\theta)^{n - \sum x_i} \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1} d\theta \\
&= \frac{B\big(a + \sum x_i,\, b + n - \sum x_i\big)}{B(a,b)},
\end{aligned}
$$

by the integral definition of the Beta function.

Example continued

Let $a_n = a + \sum x_i$ and $b_n = b + n - \sum x_i$.

It follows that the posterior distribution is $p(\theta|x_{1:n}) = \text{Beta}(\theta|a_n, b_n)$.

The posterior predictive can be derived to be

$$
\begin{aligned}
\mathbb{P}(X_{n+1} = 1 \mid x_{1:n}) &= \int \mathbb{P}(X_{n+1} = 1 \mid \theta) p(\theta|x_{1:n}) d\theta \\
&= \int \theta\ \text{Beta}(\theta|a_n, b_n) = \frac{a_n}{a_n + b_n},
\end{aligned}
$$

hence, the posterior predictive p.m.f. is

$$p(x_{n+1}|x_{1:n}) = \frac{a_n^{x_{n+1}} b_n^{1-x_{n+1}}}{a_n + b_n} \mathbb{1}(x_{n+1} \in \{0, 1\}).$$

Intro to Decision Theory

- Motivational example of DT (see my own notes for this)

General setup

Assume an unknown state S (a.k.a. the state of nature). Assume

- we receive an observation x,
- we take an action a, and
- we incur a real-valued loss $\ell(S, a)$.

S	state (unknown)
x	observation (known)
a	action
$\ell(s, a)$	loss

In the Bayesian approach,

- S is a random variable,
- the distribution of x depends on S,
- and the optimal decision is to choose an action a that minimizes the ***posterior expected loss***,

$$\rho(a, x) = \mathbb{E}(\ell(S, a)|x).$$

In other words, $\rho(a, x) = \sum_s \ell(s, a)p(s|x)$ if S is a discrete random variable, while if S is continuous then the sum is replaced by an integral.

- A ***decision procedure*** δ is a systematic way of choosing actions a based on observations x. Typically, this is a deterministic function $a = \delta(x)$ (but sometimes introducing some randomness into a can be useful).
- A ***Bayes procedure*** is a decision procedure that chooses an a minimizing the posterior expected loss $\rho(a, x)$, for each x.
- Note: Sometimes the loss is restricted to be nonnegative, to avoid certain pathologies.

Example 1

- State: $S = \boldsymbol{\theta}$
- Observation: $x = x_{1:n}$
- Action: $a = \hat{\theta}$
- Loss: $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ (quadratic loss, a.k.a. square loss)

What is the optimal decision rule?

- Goal: Minimize the posterior risk
- First note that
$$\ell(\theta, \hat{\theta}) = \theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2$$
- It then follows that
$$\rho(\hat{\theta}, x_{1:n}) = \mathbb{E}(\ell(\boldsymbol{\theta}, \hat{\theta})|x_{1:n}) = \mathbb{E}(\boldsymbol{\theta}^2|x_{1:n}) - 2\hat{\theta}\mathbb{E}(\boldsymbol{\theta}|x_{1:n}) + \hat{\theta}^2,$$
which is convex as a function of $\hat{\theta}$.

Setting the derivative with respect to $\hat{\theta}$ equal to 0, and solving, we find that the minimum occurs at $\hat{\theta} = \mathbb{E}(\boldsymbol{\theta}|x_{1:n})$, the posterior mean.

Example 2 (Perhaps have them do this in lab)

- Assume X_{n+1} is a discrete random variable.
- Setup:
  - State: $S = X_{n+1}$
  - Observation: $x = x_{1:n}$
  - Action: $a = \hat{x}_{n+1}$
  - Loss: $\ell(s, a) = \mathbb{1}(s \neq a)$ (this is called the $0-1$ loss)
- Using $0-1$ loss here works out nicely, since it turns out that the optimal decision is simply to predict the most probable value according to the posterior predictive distribution, i.e.,

$$\hat{x}_{n+1} = \delta(x_{1:n}) = \arg\max_{x_{n+1}} p(x_{n+1}|x_{1:n}).$$

Resource allocation for disease prediction

- Suppose public health officials in a small city need to decide how much resources to devote toward prevention and treatment of a certain disease, but the fraction θ of infected individuals in the city is unknown.
- Suppose they allocate enough resources to accomodate a fraction c of the population. If c is too large, there will be wasted resources, while if it is too small, preventable cases may occur and some individuals may go untreated. After deliberation, they tentatively adopt the following loss function:

$$\ell(\theta, c) = \begin{cases} |\theta - c| & \text{if } c \geq \theta \\ 10|\theta - c| & \text{if } c < \theta. \end{cases}$$

- By considering data from other similar cities, they determine a prior $p(\theta)$. For simplicity, suppose $\boldsymbol{\theta} \sim \text{Beta}(a, b)$ (i.e., $p(\theta) = \text{Beta}(\theta|a, b)$), with $a = 0.05$ and $b = 1$.
- They conduct a survey assessing the disease status of $n = 30$ individuals, $x_1, \ldots, x_n$. This is modeled as $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, which is reasonable if the individuals are uniformly sampled and the population is large. Suppose all but one are disease-free, i.e., $\sum_{i=1}^{n} x_i = 1$.

The Bayes procedure

The Bayes procedure is to minimize the posterior expected loss

$$\rho(c, x) = \mathbb{E}(\ell(\boldsymbol{\theta}, c)|x) = \int \ell(\theta, c) p(\theta|x) d\theta$$

where $x = x_{1:n}$.

- We know $p(\theta|x)$ from Equation , so we can numerically compute this integral for each c.
- Figure 1 shows $\rho(c, x)$ for our example. To visualize why it looks like this, think about the shape of $\ell(\theta, c)$ as a function of c, for some fixed θ—then imagine how it changes as θ goes from 0 to 1, and think about taking a weighted average of these functions, with weights determined by $p(\theta|x)$.
- The minimum occurs at $c \approx 0.08$, so under the assumptions above, this is the optimal amount of resources to allocate. Note that this makes more sense than naively choosing $c = \bar{x} = 1/30 \approx 0.03$, which does not account for uncertainty in θ and the large loss that would result from possible under-resourcing.

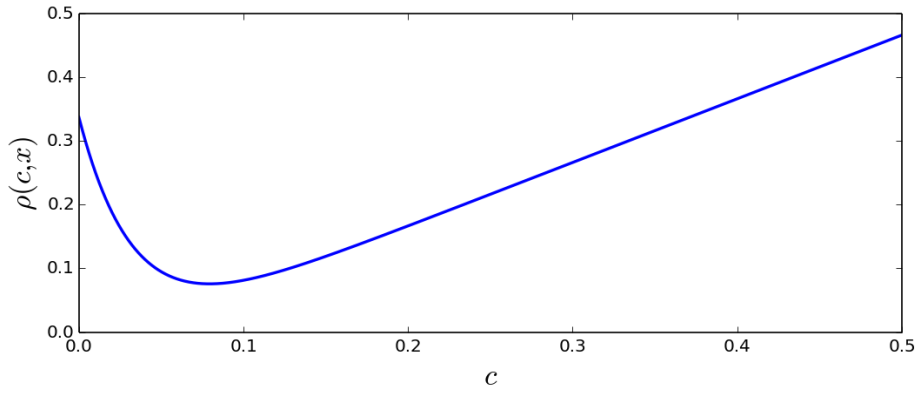


Figure 1: Posterior expected loss for the disease prevalence example.

- (Note: A sensitivity analysis should also be performed to assess how much these results depend on the assumptions. More on this later.)