COLUMBIA UNIVERSITY

IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Binomial Model

Assume we have binary data

$$X_1, X_2, \dots, X_n | \, \theta \ \sim \text{Bernoulli}(\theta)$$

Note that $f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}, i = 1, \dots, n$

Recall from last time the conditional likelihood is:

$$\begin{aligned} \mathrm{L}(\theta) = f(x_1, \dots, x_n|\theta) &= \prod_{i=1}^{n} [\theta^{x_i}(1-\theta)^{1-x_i}] \\ &= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{\sum_{i=1}^{n}(1-x_i)} \\ &= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} \end{aligned}$$

What remains to be specified is the prior distribution.

Uniform Prior

Suppose we have no information about the value of θ. This means we assume

$$\theta \sim \mathrm{U}(0, 1)$$

$$\pi(\theta) = 1, \qquad 0 \leq \theta \leq 1$$

Then the posterior is

$$f(\theta|x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n|\theta)\pi(\theta)}{p(x_1, \dots, x_n)}$$

$$= p(x_1, \dots, x_n|\theta) \times \frac{1}{p(x_1, \dots, x_n)} \propto p(x_1, \dots, x_n|\theta)$$

The last line says that $f(\theta|x_1, \dots, x_n)$ and $p(x_1, \dots, x_n|\theta)$ are proportional to each other as functions of θ.

Example 1

Suppose you are playing a new game and want to estimate your chance θ of winning it. You don't know anything about the game before you start playing so the prior is assumed uniform. Then you play 9 games, and they result in the following sequence:

$$W\ L\ W\ W\ W\ L\ W\ L\ W$$

Plot the posterior distribution of your chance of winning using a grid approximation.

Example 1 Solution

You play 9 games, and they result in the following sequence:

$$W\,L\,W\,W\,W\,L\,W\,L\,W$$

Plot the posterior distribution of your chance of winning using a grid approximation.

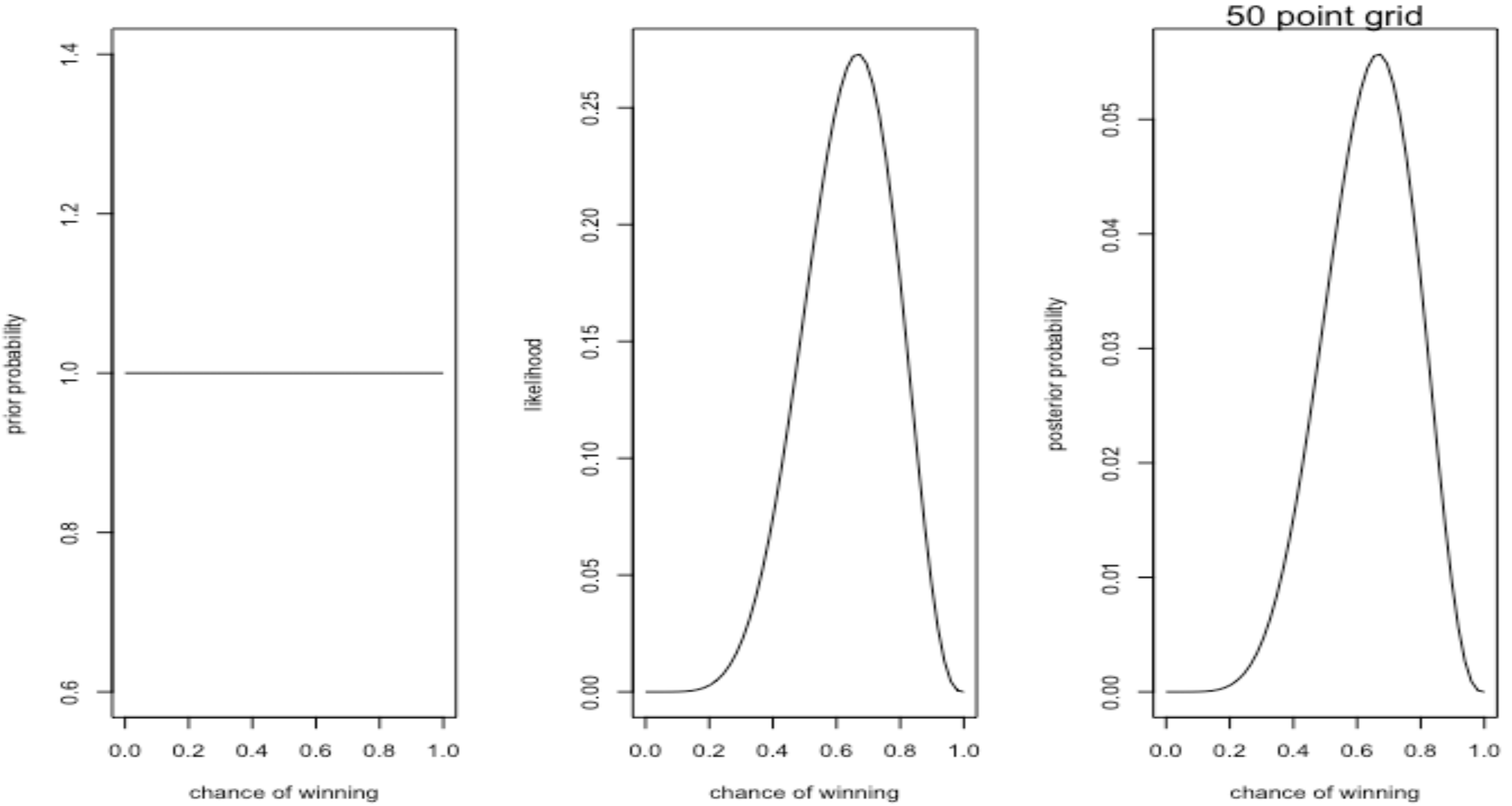
We have $n = 9$, and $x_1 = 1, x_2 = 0, \dots, x_9 = 1, \sum_{i=1}^{9} x_i = 6$

and the posterior is

$$f(\theta|x_1, \dots, x_9) \propto p(x_1, \dots, x_n|\theta) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$
$$= \theta^6\,(1-\theta)^3$$

Notice this is proportional to the Binomial probability of 6 successes out of 9 trials.

Example 1 Graph



Example 1 Continued

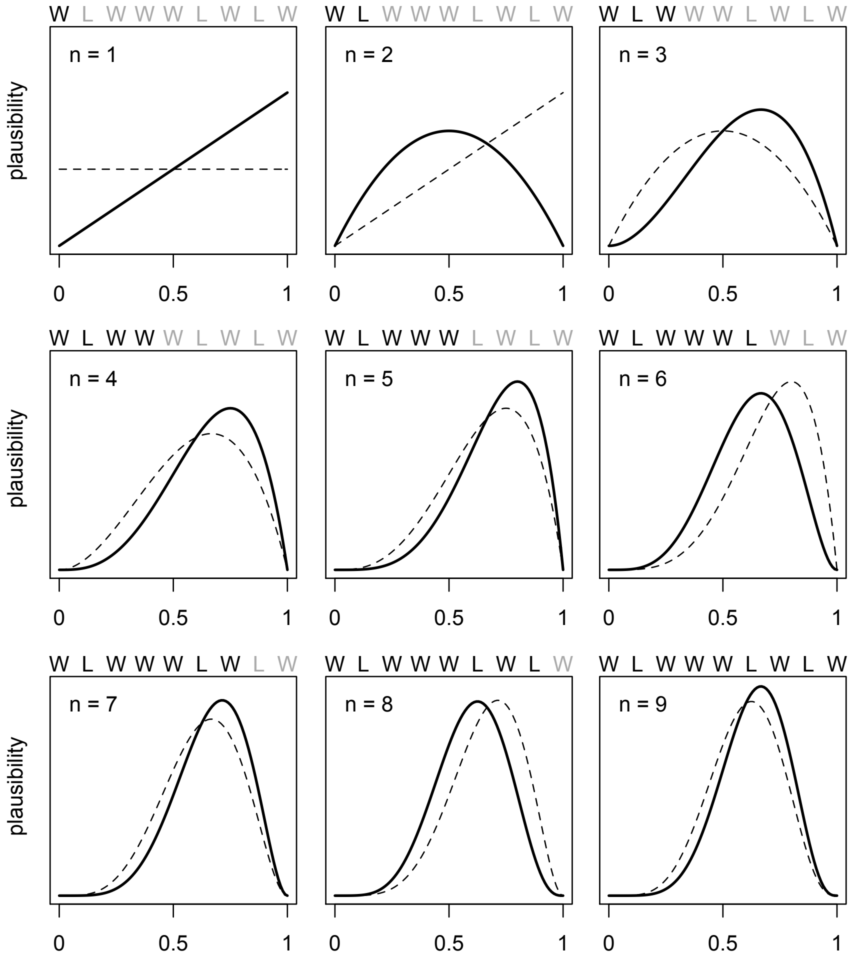
Note that if we use the data sequentially

$$W\ L\ W\ W\ W\ L\ W\ L\ W$$

and at each step the posterior of the previous step is the prior for the next, then at the end the final graph is the same as what we would get if we use the entire sample with the original prior.

This is a well-known phenomenon in every Bayesian analysis.

Example 1 Sequential Graph



Exercise 1

Repeat the Example 1 with the same data, but this time using a prior which reflects your confidence that the chance of winning the game is definitely in your favor, but you don’t know anything else about it.

Normalizing Constant

To find the marginal density of the data, observe the following:

$$p(x_1, \dots, x_n) = \int_0^1 p(x_1, \dots, x_n | \theta) \pi(\theta) d\theta$$

$$= \int_0^1 \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} \, d\theta$$

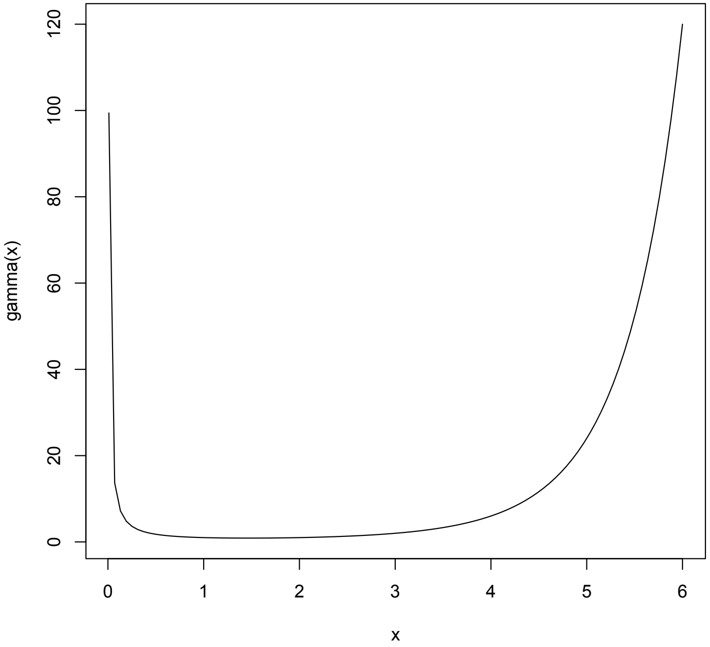
Recall from calculus:

$$\int_0^1 \theta^{a-1} (1 - \theta)^{b-1} d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \forall a, b > 0$$

Therefore,

$$p(x_1, \dots, x_n) = \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}$$

where $x = \sum_{i=1}^n x_i$



Aside: Gamma Function

- The Gamma function is defined for $x > 0$ by

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$$

- It is available in R with the `gamma` function.

Example 2

Prove that

$$\Gamma(x + 1) = x\Gamma(x)$$

Example 2 Solution

Proof:

$$\Gamma(x+1)=\int_{0}^{\infty} u^{x}e^{-u}du$$

We will use integration by parts to evaluate the integral.

Define $w = u^x$ and $v = -e^{-u}$. Then $dv = e^{-u}du$ and $dw = xu^{x-1}dx$

Therefore,

$$\int_{0}^{\infty} u^{x}e^{-u}du = -u^{x}e^{-u}]_{u=0}^{u=\infty} + x\int_{0}^{\infty} u^{x-1}e^{-u}du$$

$$= 0 + x\Gamma(x)$$

Corollary: $\Gamma(n)=(n-1)!, \forall\, n \in \mathbb{N}$

Exercise 2

Prove that

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Posterior Distribution

We can now find the posterior distribution exactly:

$$f(\theta|x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n|\theta)\pi(\theta)}{p(x_1, \dots, x_n)}$$

$$= \frac{\theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}}{\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}}$$

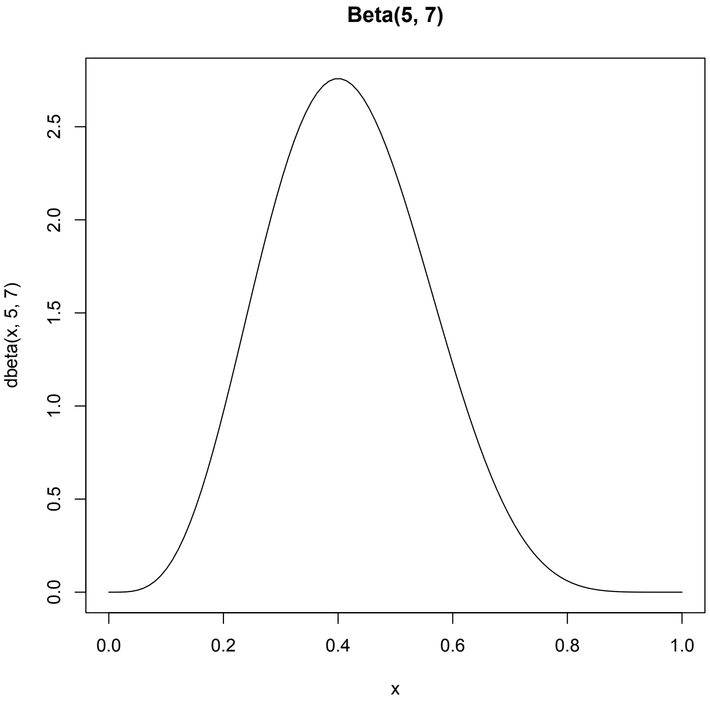
$$= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^{x}(1-\theta)^{n-x}, \theta \in (0,1)$$

where $x = \sum_{i=1}^{n} x_i$

It can even be recognized that

$$\theta|x_1, \dots, x_n \sim \mathrm{Beta}(x+1, n-x+1)$$

That is, the posterior has beta distribution with parameters x + 1 and n – x +1.



Aside: Beta Distribution

- The pdf of Beta(a, b) is defined for $a, b > 0$ by

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, x \in [0,1]$$

- It is available in R with the `dbeta` function.

Example 3

Let the $X \sim$ Beta(a, b)

If $a > 1$ and $b > 1$, prove that the mode of the distribution is

$$\text{mode} = \frac{a-1}{a+b-2}$$

Example 3 Solution

The mode is the max of the pdf.

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

$$\Rightarrow f'(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}[(a-1)x^{a-2}(1-x)^{b-1} - (b-1)x^{a-1}(1-x)^{b-2}]$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-2}(1-x)^{b-2}[(a-1)(1-x) - (b-1)x] = 0$$

$$\Leftrightarrow (a-1)(1-x) - (b-1)x = 0 \Leftrightarrow a-1 = (a-1+b-1)x$$

Therefore,

$$f'(x) = 0 \Leftrightarrow x = \frac{a-1}{a+b-2}$$

It can be shown that when $a > 1$ and $b > 1$ the second derivative test confirms that this is the max of the pdf.

Note: when $0 < a < 1$ and $0 < b < 1$ the mode is at either 0 or 1 and the pdf is convex.

Exercise 3

Let the $X \sim$ Beta(a, b)

Prove that:

$$\mathrm{E}(X) = \frac{a}{a+b}$$

$$\mathrm{Var}(X) = \frac{ab}{(a+b+1)(a+b)^2}$$

Different Prior

Notice that the uniform distribution we used is equivalent to beta distribution with parameters $a = b = 1$. That is,

$$\mathrm{U}(0, 1) = \mathrm{Beta}(1, 1)$$

Suppose now that the prior is any beta distribution:

$$\theta \sim \mathrm{Beta}(a, b)$$

Then the posterior is

$$f(\theta|x_1, \dots, x_n) =$$

$$= \frac{\theta^x (1-\theta)^{n-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}}{p(x_1, \dots, x_n)}$$

$$\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1}, \theta \in (0,1)$$

where $x = \sum_{i=1}^{n} x_i$

The only distribution with such shape is Beta($a + x$, $b + n - x$).

Throughout the course we will use this trick to identify posterior distributions.

Conjugate Priors

We have shown that a beta prior distribution and a binomial sampling model led to a beta posterior distribution, that is, the prior and posterior are of the same “type”. In the past computational drawbacks calculating the posterior used to be so severe that many researchers worked only with the so-called *conjugate* priors.

Definition: A family of probability distributions $\mathcal{F} = \{\, \pi(\theta) \,\}$ is conjugate for the model $f(x \mid \theta)$ if

$$f(\theta \mid x) \in \mathcal{F}$$

Note: The Gaussian example from Lecture 1 was also conjugate, because the prior had normal distribution and the posterior also had a normal distribution.

Posterior Analysis

In the Beta-Binomial model with Beta(a, b) prior we showed the posterior is:

$$f(\theta|x_1, \dots, x_n) = \text{Beta}(a + x, b + n - x)$$

Using the results from Example 3 and Exercise 3 we obtain:

$$\mathrm{E}(\theta|x) = \frac{a + x}{a + b + n}$$

$$\mathrm{Var}(\theta|x) = \frac{(a + x)(b + n - x)}{(a + b + n + 1)(a + b + n)^2}$$

$$\mathrm{mode}(\theta|x) = \frac{a + x - 1}{a + b + n - 2}$$

All of these are useful posterior summaries and we will analyze them more closely.

Posterior Mean

We have that:

$$\mathrm{E}(\theta|x) = \frac{a+x}{a+b+n} = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{x}{n}$$

$$= \frac{a+b}{a+b+n}\times\text{prior expextation} + \frac{n}{a+b+n}\times\text{data average}$$

That is, for this model and prior distribution, the posterior expectation (also known as the posterior mean) is a weighted average of the prior expectation and the sample average, with weights proportional to $a + b$ and n respectively.

If our sample size n is larger than $a + b$, then it seems reasonable that a majority of our information about θ should be coming from the data as opposed to the prior distribution.

Indeed, if $n >> a + b$, then $\mathrm{E}(\theta|x) \approx \frac{x}{n}$

Prediction

An important feature of Bayesian inference is the existence of a predictive distribution for a new observation x_{new}. It is defined as the conditional distribution of x_{new} given $\{x_1, \ldots, x_n\}$.

$$P(X_{\text{new}} = 1|x_1, \ldots, x_n) = \int_0^1 f(X_{\text{new}} = 1, \theta|x_1, \ldots, x_n)d\theta$$

$$= \int_0^1 f(X_{\text{new}} = 1|\theta, x_1, \ldots, x_n)f(\theta|x_1, \ldots, x_n)d\theta$$

$$= \int_0^1 \theta f(\theta|x_1, \ldots, x_n)d\theta = E(\theta|x) = \frac{a + \sum_{i=1}^n x_i}{a + b + n}$$

Note: The predictive distribution does not depend on any unknown quantities. If it did, we would not be able to use it to make predictions.

Example 4

Suppose again that the prior is the uniform distribution Beta(1, 1). Then

$$P(X_{\text{new}} = 1|x_1, \dots, x_n) = \frac{1 + \sum_{i=1}^{n} x_i}{1 + 1 + n} = \frac{x + 1}{n + 2}$$

where $x = \sum_{i=1}^{n} x_i$ is the total number of successes in the sample.

Note that this is the same as the so called "Wilson correction" to the traditional MLE estimate $P(X_{\text{new}} = 1|x_1, \dots, x_n) = \frac{x}{n}$. It was derived without reference to Bayesian analysis to avoid situations where $x = 0$ or $x = n$.

Exercise 4

Derive the variance of the predictive distribution for the Beta-Binomial model with Beta(a, b) prior.

Confidence Regions

It is often desirable to identify regions of the parameter space that are likely to contain the true value of the parameter.

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

$$P[l(x) < \theta < u(x)|X = x)] = 0.95.$$

The interpretation of this interval is that it describes your information about the location of the true value of θ after you have observed $X = x$.

In contrast, the frequentist CI has the property that when you plug in the data into the formula, it covers θ with probability either 0 or 1. This highlights the lack of a post-experimental interpretation of frequentist coverage.

Quantiles Based Interval

Perhaps the easiest way to obtain a confidence interval is to use posterior quantiles. To make a $100 \times (1 - \alpha)\%$ quantile-based confidence interval, find numbers $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ such that:

1. $P(\theta < \theta_{\alpha} | X = x) = \alpha/2$;
2. $P(\theta > \theta_{1-\alpha/2} | X = x) = \alpha/2$.

That is, the numbers $\theta_{\alpha/2}$ and $\theta_{1-\alpha/2}$ are the posterior $\alpha/2$ and $1 - \alpha/2$ quantiles computed from the posterior distribution.

Note: this is not the best approach, and in fact, the interval can be chosen such that the interval still has the same coverage but is often of narrower width.

Example 5

Suppose we have Bernoulli data with sample size n = 9 and total number of successes x = 1. Suppose further we utilized the U(0, 1) prior. Then the posterior distribution is:

$$\theta \mid X = 1 \sim \text{Beta}(1 + 1, 8 \ + 1) = \text{Beta}(2, 9)$$

You can't compute the 0.025 and 0.975 quantiles manually, so you need to use the R function `qbeta`.

```
qbeta( c(.025 ,.975) , 2, 9)
0.02521073 0.44501612
```

That is, the probability that $\theta \in [0.02521, 0.0445] = 0.95$.

Exercise 5

Compute an 80% quantile CI if we have n = 17 observation with a total of 14 successes and prior Beta(8, 2).

Highest Posterior Density Region

Definition: A 100 × (1 − α)% *highest posterior density* (HPD) region consists of a subset of the parameter space, $s(x) \subset \Theta$ such that:

1. $P[\theta \in s(x) \mid X = x] = 1 - \alpha$
2. If $\theta_a \in s(x)$, and $\theta_b \notin s(x)$, then $f(\theta_a|X = x) > f(\theta_B|X = x)$

Note this means that all points in an HPD region have a higher posterior density than points outside the region. However, an HPD region might not be an interval if the posterior density is multimodal (having multiple peaks).

Confidence Regions

The HPD regions are computed by finding:

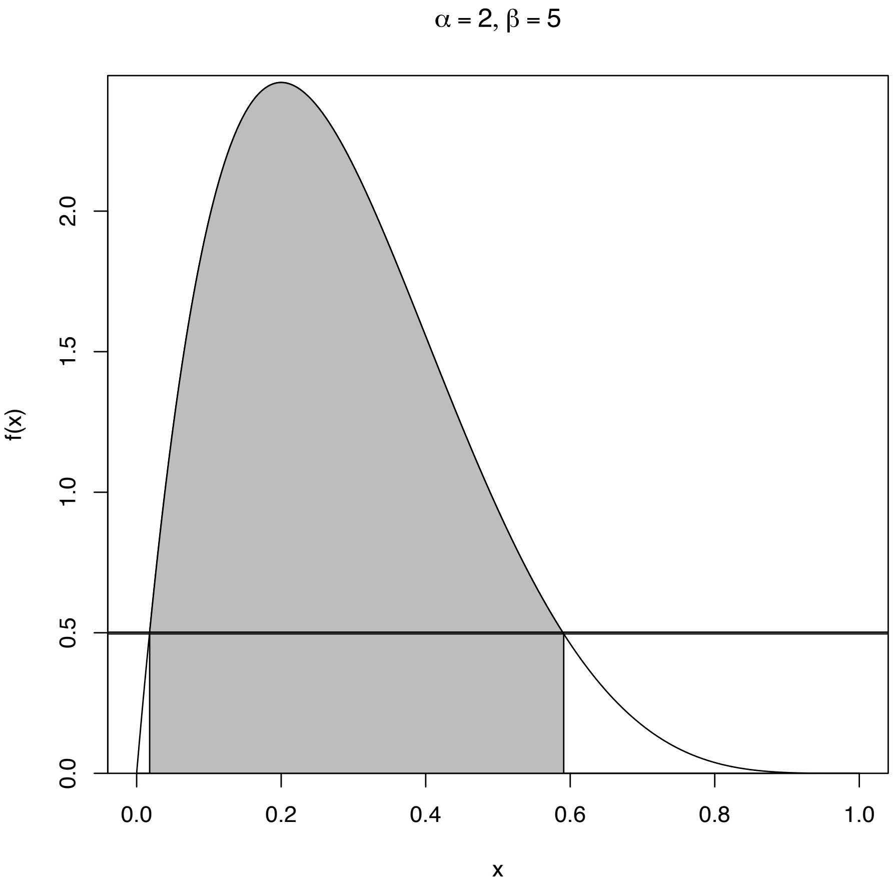
$$C(\boldsymbol{x}) = \{\theta : f(\theta \mid \boldsymbol{x}) \geq k\}$$

where k is such that

$$P(\theta \in C(\boldsymbol{x}) \mid \boldsymbol{x}) = \gamma$$
$$= P\left(f(\theta \mid \boldsymbol{x}) \geq k \mid \boldsymbol{x}\right)$$

This leads to a different computational problem, namely solving

$$f(\theta \mid \boldsymbol{x}) = k$$


α = 2, β = 5
f(x)
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
x