Bayesian Inference

Bret Larget

May 9, 2014

Abstract

The Bayesian approach to statistics is based on an alternative philosophy that treats parameters (and anything uncertain) as random. The approach requires a prior probability distribution for each unknown parameter whose distribution is updated as a posterior distribution upon receiving more information, as governed by Bayes' theorem.

1 The Bayesian Approach

In the Bayesian approach to statistics, anything unknown (like the truth of hypotheses or the values of parameters) is treated as random, and as such, is described by a probability distribution. The Bayesian philosophical approach to probability is different than that of the frequentist philosophy (which is central to the methods in the course up to this point). From the frequentist point of view, parameters are fixed and unknown; hypotheses are true or false. As these objects are not random, the frequentist says uncertainty about them cannot be described with probability. The Bayesian point of view is that all uncertainty must be described by probability.

1.1 Prior and Posterior Distributions, Likelihood

Before seeing data, the *prior distribution* of an unknown parameter θ is described by a probability density (or mass function, if discrete) $f(\theta)$.[1] The Bayesian approach connects data and parameter through the *likelihood function*, $f(x \mid \theta)$. Recall that when treating the parameter θ as fixed, integrating (or summing) over all possible data values x results in a total one. But the same function where x is fixed as the data and the parameter θ is what varies is called the likelihood. Parameter values where the likelihood is high are those that have a high probability of producing the observed data. In the maximum likelihood approach to statistics, the best estimate of the value $\hat{\theta}$ that maximizes the likelihood (and log-likelihood) function. All Bayesian inference is based on evaluation of the *posterior distribution*, which,

[1] In the description that follows, the symbol f will be used to describe multiple different functions; each is distinguished by its argument, so $f(\theta)$ and $f(x)$ will mean different probability distributions, not the same function evaluated at two different values.

by Bayes' Theorem, is written as

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)}$$

where the only term we have not defined is the *marginal likelihood* $f(x)$, which is the probability of observing the data x averaged across the entire parameter space.

$$f(x) = \int f(x \mid \theta) f(\theta) \, \mathrm{d}\theta$$

Notice that the posterior density is proportional to the product of the likelihood and the prior density; the marginal likelihood, which is constant when the data x is observed, is merely the required constant to normalize the product of the likelihood and prior density so that when treating the product as a function of θ with x fixed, it integrates to one as is required for the left-hand-side of the equation to be a proper probability density for θ.

In the Bayesian approach to statistical inference, the best values of θ for explaining the observed data x are found by combining information from the prior distribution (which values of θ are probable based on our prior understanding of the setting) and the likelihood (which values of θ are likely to produce the observed data). Note as well that different choices of prior distribution can result in different inference from the same data. Critics of the Bayesian approach to statistical inference focus on the subjective nature of selecting a prior distribution in the absence of data. Proponents of the Bayesian approach see the direct probabilistic interpretation of inference to be advantageous and see the selection of a prior distribution as either a natural way to incorporate true prior information into the inference process, as a way to specify an individuals subjective prior beliefs, or as a pragmatic concern where choosing an appropriately vague prior distribution will have very small impact on the final inference given informative data.

2 Theory for a Single Population Proportion

To make the previous section concrete, consider an example of estimating an unknown population parameter p, say the proportion of red balls in an infinitely large population, based on a sample of n individuals where there are x red balls in the sample. The natural likelihood function for the setting is the binomial distribution.

$$f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad 0 < p < 1$$

In theory, the prior distribution for p could be discrete or continuous; it can be any legitimate probability distribution for valid values of p between 0 and 1. For reasons of mathematical convenience, we will focus on a continuous prior probability distribution called the *Beta distribution* which is determined by two parameters α and β and has density $f(p)$ proportional to $p^{\alpha-1}(1-p)^{\beta-1}$. This product needs to be divided by the correct normalizing constant so that the integral of the probability density is one; namely

$$\int_0^1 C p^{\alpha-1} (1-p)^{\beta-1} \, \mathrm{d}p = 1$$

so that we can find C by solving this problem.

$$\frac{1}{C} = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\,\mathrm{d}p$$

It is required that $\alpha > 0$ and $\beta > 0$ or the previous integral will not be finite. The integral on the right-hand-side of the above equation is called the *beta function*, $B(\alpha, \beta)$, and has solution

$$B(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\,\mathrm{d}p = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

where the *gamma function*, Γ, is defined as

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

This looks complicated, but is related to things we have seen before. For a positive integer n, $\Gamma(n) = (n-1)!$, so the gamma function is a continuous interpolation of the factorial function. The following can be shown by direct mathematical arguments.

$$\begin{aligned} \Gamma(\alpha+1) &= \alpha\Gamma(\alpha) \\ \Gamma(1) &= 1 \\ \Gamma\left(\frac{1}{2}\right) &= \sqrt{\pi} \end{aligned}$$

All of this is simply theory to specify a convenient choice of prior density for p, where $p \sim \text{Beta}(\alpha_0, \beta_0)$.

$$f(p) = \left(\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\right) p^{\alpha_0-1}(1-p)^{\beta_0-1}, \qquad 0 < p < 1$$

Conjugate Prior Distribution

Note that both the likelihood $f(x \mid p)$ and prior density $f(p)$ are constants multiplied by a function of p of the form p to a power times $(1-p)$ to a power. The posterior density is then proportional to the product of likelihood and prior density.

$$f(p \mid x) \propto p^x(1-p)^{n-x} \times p^{\alpha_0-1}(1-p)^{\beta_0-1} = p^{(x+\alpha_0)-1}(1-p)^{(n-x+\beta_0)-1}$$

The posterior density indicates that the posterior distribution of p also has a Beta distribution, but with $\alpha = x + \alpha_0$ and $\beta = n - x + \beta_0$. To summarize, if

$$\begin{aligned} p &\sim \text{Beta}(\alpha_0, \beta_0) \qquad \text{and} \\ X \mid p &\sim \text{Binomial}(n, p) \end{aligned}$$

then

$$p \mid X \sim \text{Beta}(x + \alpha_0, n - x + \beta_0)$$

As more data is gathered, the parameter α increases with each success and β increases with each failure.

Moments of the Beta Distribution

The first two moments of the beta distribution are:

$$\text{mean} = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \text{variance} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Notice that the variance may also be written in this fashion.

$$\text{variance} = \left(\frac{\alpha}{\alpha+\beta}\right)\left(\frac{\beta}{\alpha+\beta}\right)\left(\frac{1}{\alpha+\beta+1}\right)$$

Notice that the variance gets smaller as the sum $\alpha+\beta$ increases. When applied to the posterior distribution, this is

$$\text{variance} = \left(\frac{x+\alpha_0}{n+\alpha_0+\beta_0}\right)\left(\frac{n-x+\beta_0}{n+\alpha_0+\beta_0}\right)\left(\frac{1}{n+\alpha_0+\beta_0+1}\right)$$

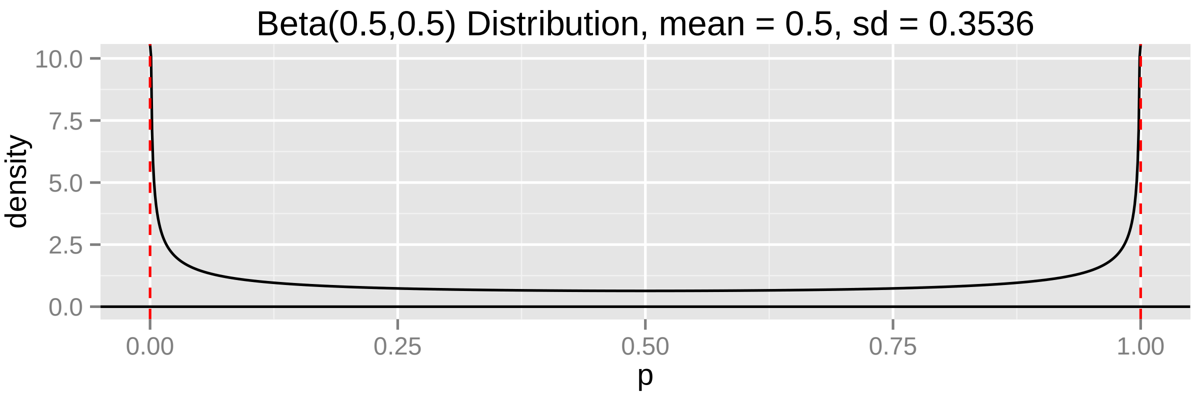
which, in the case where $\alpha_0 = \beta_0 = 2$ and $\tilde{p} = (x+2)/(n+4)$ simplifies to

$$\text{variance} = \frac{\tilde{p}(1-\tilde{p})}{n+5}$$

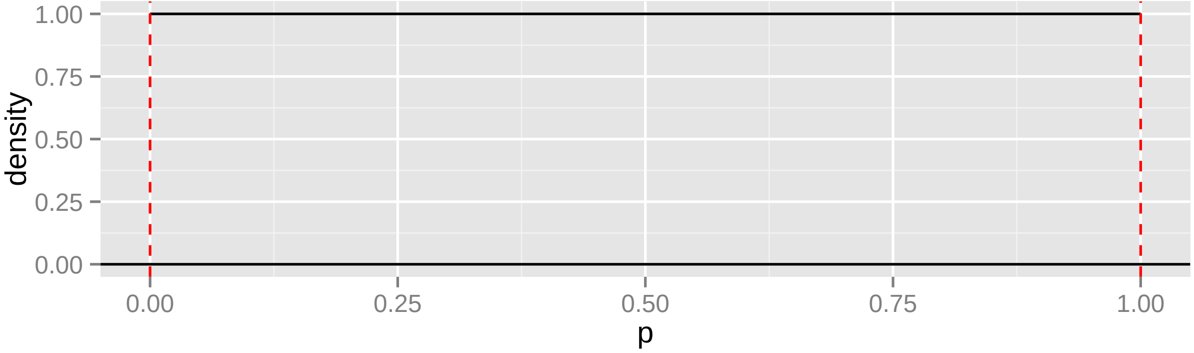
which is very close to the method for confidence intervals for p we studied earlier in the semester. The only difference is that the denominator is $n+5$ rather than $n+4$. This method of inference is quite close in practice to the Bayesian approach for the prior with $\alpha_0 = \beta_0 = 2$.

3 Pictures

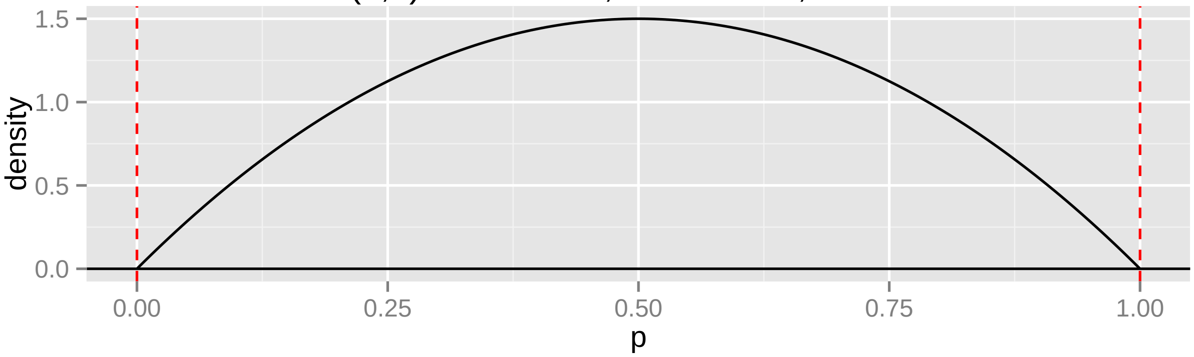
We are likely well past the point where a few pictures to show what is going on would be helpful! First, here are multiple examples of different prior densities. Note that when α_0 and β_0 are each larger than one, the density goes to zero and $p = 0$ and $p = 1$ but that it goes to infinity when $\alpha_0 < 1$ and $\beta_0 < 1$. When $\alpha + 0 = \beta_0 = 1$, the distribution is uniform. When $\alpha_0 + \beta_0$ is larger, the prior distribution is more concentrated. Thus, good choices typically have $\alpha_0 + \beta_0$ fairly small relative to the sample size so that the likelihood and data will dominate the posterior distribution.



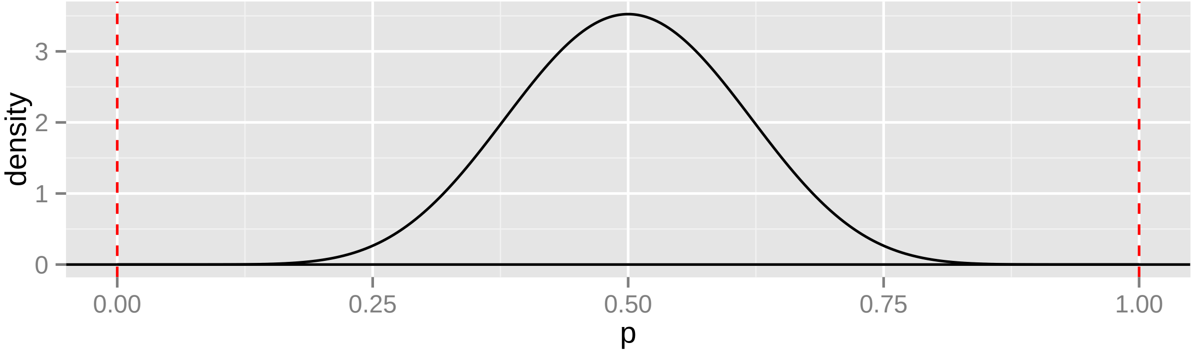
Beta(1,1) Distribution, mean = 0.5, sd = 0.2887

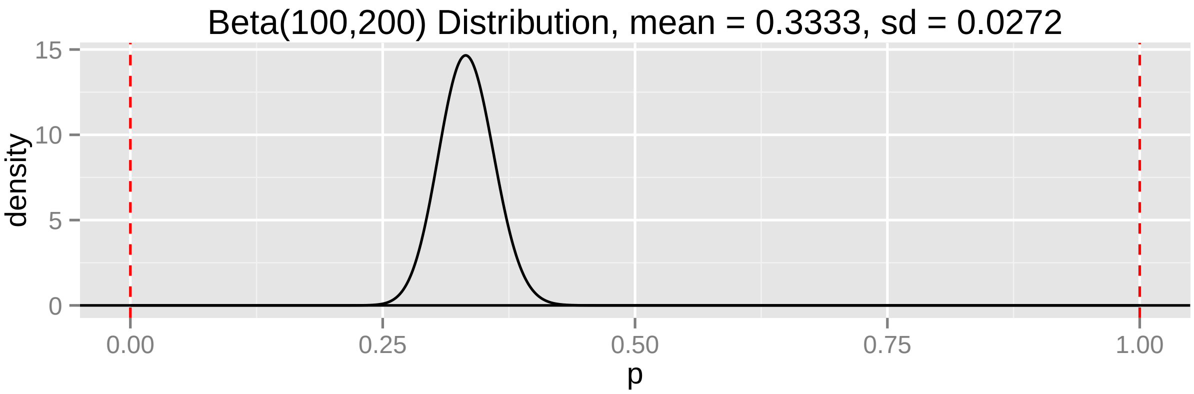


Beta(2,2) Distribution, mean = 0.5, sd = 0.2236



Beta(10,10) Distribution, mean = 0.5, sd = 0.1091

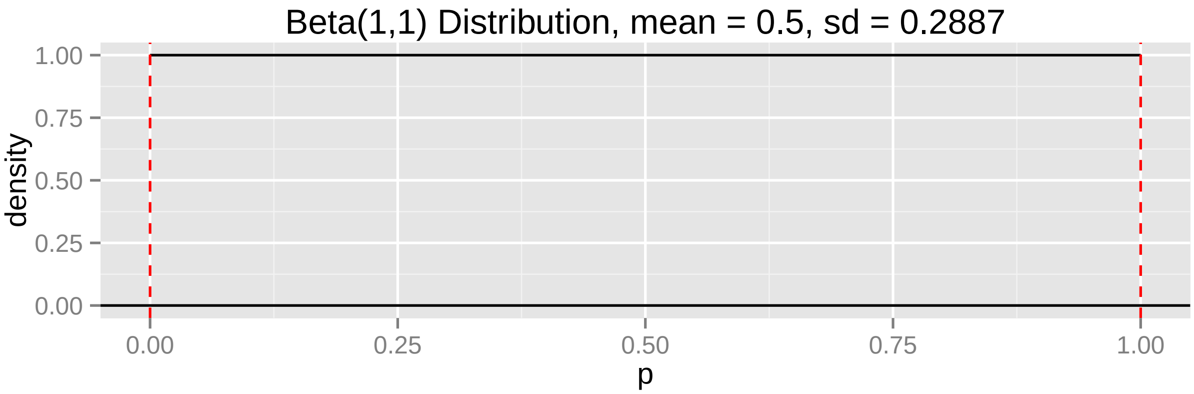




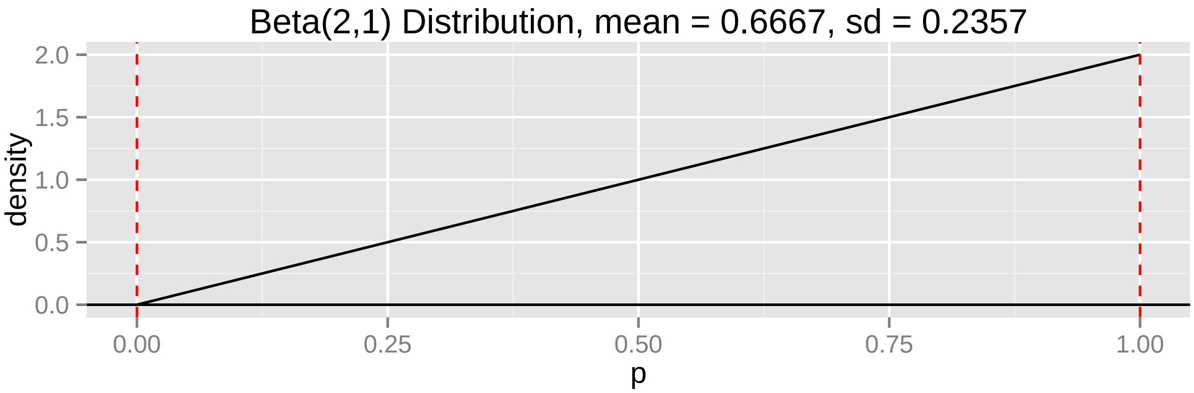
4 Inference

Let's see what happens to the posterior distribution of p as we get new data, beginning with a uniform prior density.

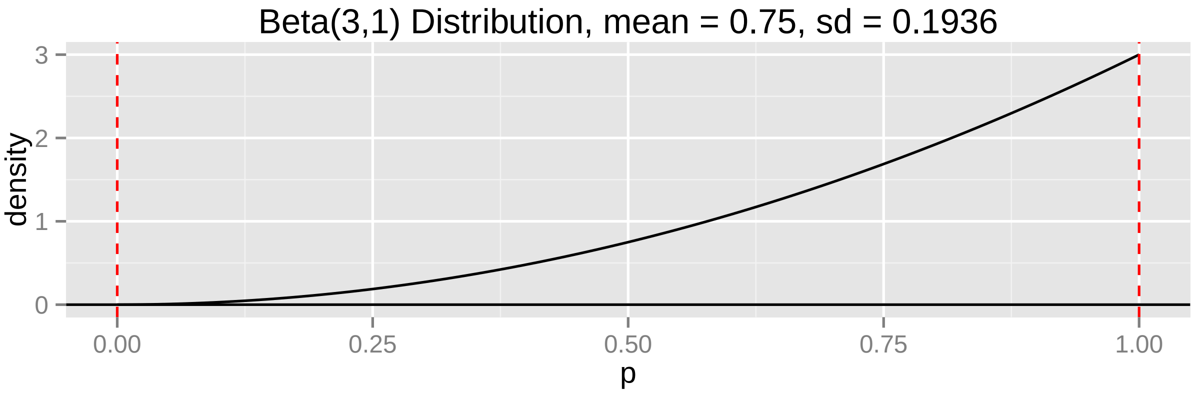
Before we see data, here is a graph of the posterior distribution.



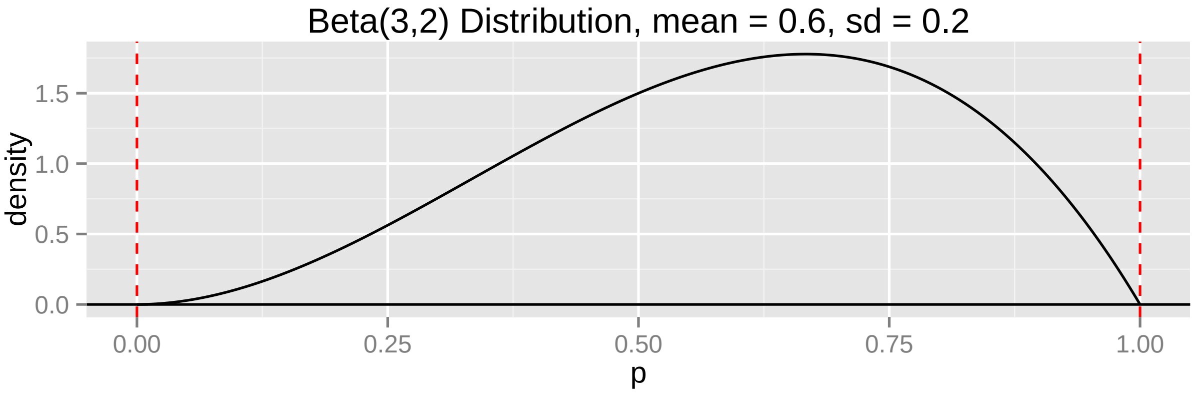
Suppose the first draw is a red ball. We now think that p, the proportion of red balls, is probably bigger than what we saw earlier, given this new data.



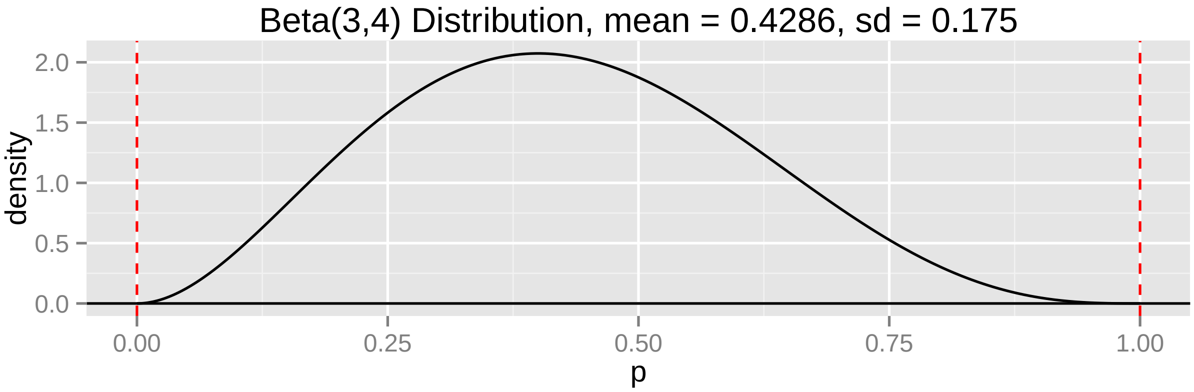
A second red ball shifts the distribution even closer to 1. But $n = 2$ is a small sample size, so p values near zero are still not too improbable.



A white ball shifts the distribution back toward 0.



Two more white balls shifts it further.



Notice that as we get more data, the distribution gets more concentrated (although slowly) and that the shape is beginning to resemble a normal curve (at least as the number of red and white balls both increase).

Also note here that the mean of the beta distribution is $3/7 \doteq 0.4286$. We have seen $2/5 = 0.4$ red balls so far, but the posterior density is centered somewhere between 0.4 and

0.5, which was the mean of the uniform prior density. This is not a coincidence.

Posterior Mean

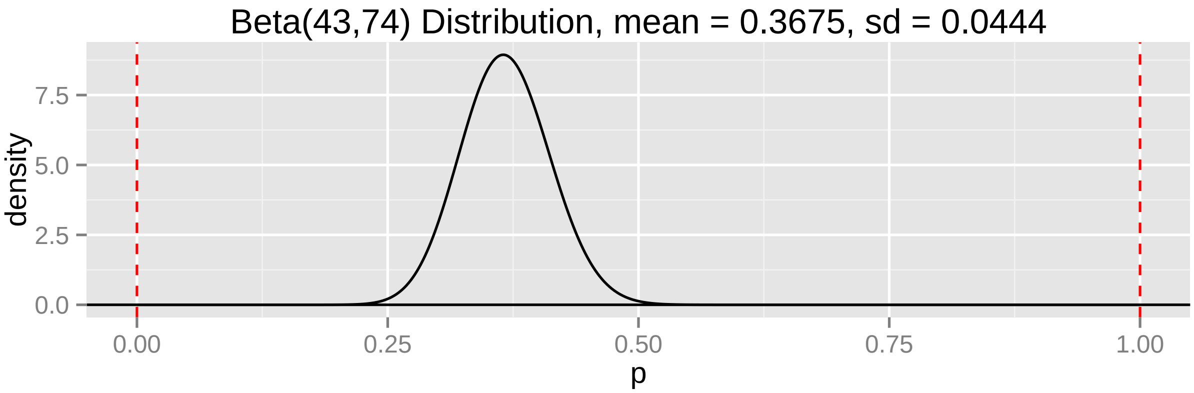
The posterior mean is

$$\mathsf{E}(p \mid x) = \frac{x + \alpha_0}{n + \alpha_0 + \beta_0} = \left(\frac{\alpha_0 + \beta_0}{n + \alpha_0 + \beta_0}\right)\left(\frac{\alpha_0}{\alpha_0 + \beta_0}\right) + \left(\frac{n}{n + \alpha_0 + \beta_0}\right)\left(\frac{x}{n}\right)$$

which means that the posterior mean is a weighted average between the prior mean and the maximum likelihood estimate where the relative weight of the prior distribution is $\alpha_0 + \beta_0$ and the relative weight of the maximum likelihood estimate is n. Hence, one can think of the sum $\alpha_0 + \beta_0$ as the effective number of observations that the prior distribution is equivalent to. If this is much smaller than n, then the data will dominate the posterior inference.

Back to Pictures

Let's jump ahead and add 40 more red balls and 70 more white balls to the sample. Notice how much more concentrated the posterior density is. After observing 42 red balls and 73 white balls, we have a pretty good idea where p is.



Credible Regions

Similar to finding a confidence interval from a bootstrap density by cutting off a fraction of the distribution on both sides, we can do the same for the Bayesian posterior distribution to find a *credible region*. To find a 95% credible region for the data so far, as simply need to find the 0.025 and 0.975 quantiles of the Beta$(43, 74)$ distribution.

```
qbeta(c(0.025, 0.975), 43, 74)

## [1] 0.2829 0.4565
```

Compare this to the standard interval using $\hat{p} = 42/115$.

```
p.hat = 42/115
se = sqrt(p.hat * (1 - p.hat)/115)
p.hat + 1.96 * c(-se, se)

## [1] 0.2772 0.4532
```

The confidence interval using $\tilde{p}$ and the Bayesian credible region when $\alpha_0 = \beta_0 = 2$ are even more similar from the same data.

```
qbeta(c(0.025, 0.975), 44, 75)

## [1] 0.2856 0.4580

p.tilde = 44/119
se.tilde = sqrt(p.tilde * (1 - p.tilde)/119)
p.tilde + 1.96 * c(-se.tilde, se.tilde)

## [1] 0.2830 0.4565
```

5 Summary

The Bayesian approach to inference differs from the frequentist approach in that probabilities are used directly to quantify anything that is uncertain. Parameters are random variables.

The posterior density is proportional to the product of the likelihood and prior density.

For inference about a single population proportion, the Bayesian approach to estimation is to find the posterior density and then cut off a given percentage on each end to state that there is, say, a 95% probability that the unknown p is in the given interval.

A confidence interval from the same data will have slightly different end points, but the practical difference between the two approaches gets small as the sample size increases, at least if a somewhat vague prior density is used for the Bayesian approach.