Beta Prior for Bernoulli

Tom Fletcher

January 21, 2016

In this note we will look at the conjugate prior of the Bernoulli distribution, which is a Beta distribution. Recall that Bernoulli is the distribution for a binary random variable. The notation $X \sim \mathrm{Ber}(\theta)$ means $P(X = 1) = \theta$ and $P(X = 0) = 1 - \theta$. Now, if we are given n realizations of this Bernoulli variable, the likelihood function for θ is

$$p(x_1, \ldots, x_n \,|\, \theta) = \prod_{i=1}^{n} P(X = x_i \,|\, \theta) = \prod_{i=1}^{n} \theta^{x_i}(1 - \theta)^{1-x_i}.$$

If we let k be the number of x_i with value 1, i.e., $k = \sum_i x_i$, the likelihood function is written in more simple form as

$$p(x_1, \ldots, x_n \,|\, \theta) = \theta^k (1 - \theta)^{n-k}.$$

Now let's look at putting a Beta distribution on the parameter θ of the Bernoulli likelihood. The notation for the Beta distribution is $\theta \sim \mathrm{Beta}(\alpha, \beta)$, and its pdf is given by

$$p(\theta) = \frac{1}{\mathrm{B}(\alpha, \beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}, \qquad \text{for } \theta \in [0, 1],$$

where $\mathrm{B}(\alpha, \beta)$ is the Beta function. For more information on the Beta distribution, see the Wikipedia page: https://en.wikipedia.org/wiki/Beta_distribution.

Notice that for $\alpha = 1, \beta = 1$ this prior reduces to the uniform distribution on $[0, 1]$. Using Bayes' rule, the posterior distribution for θ is

$$p(\theta \,|\, x_1, \ldots, x_n) = \frac{1}{C} p(x_1, \ldots, x_n \,|\, \theta) p(\theta) = \frac{1}{C} \theta^{\alpha+k-1}(1 - \theta)^{\beta+n-k-1},$$

where $C = \int p(x_1, \ldots, x_n, \theta) d\theta$ is the normalizing constant. But notice that this posterior is also a Beta distribution, with new parameters:

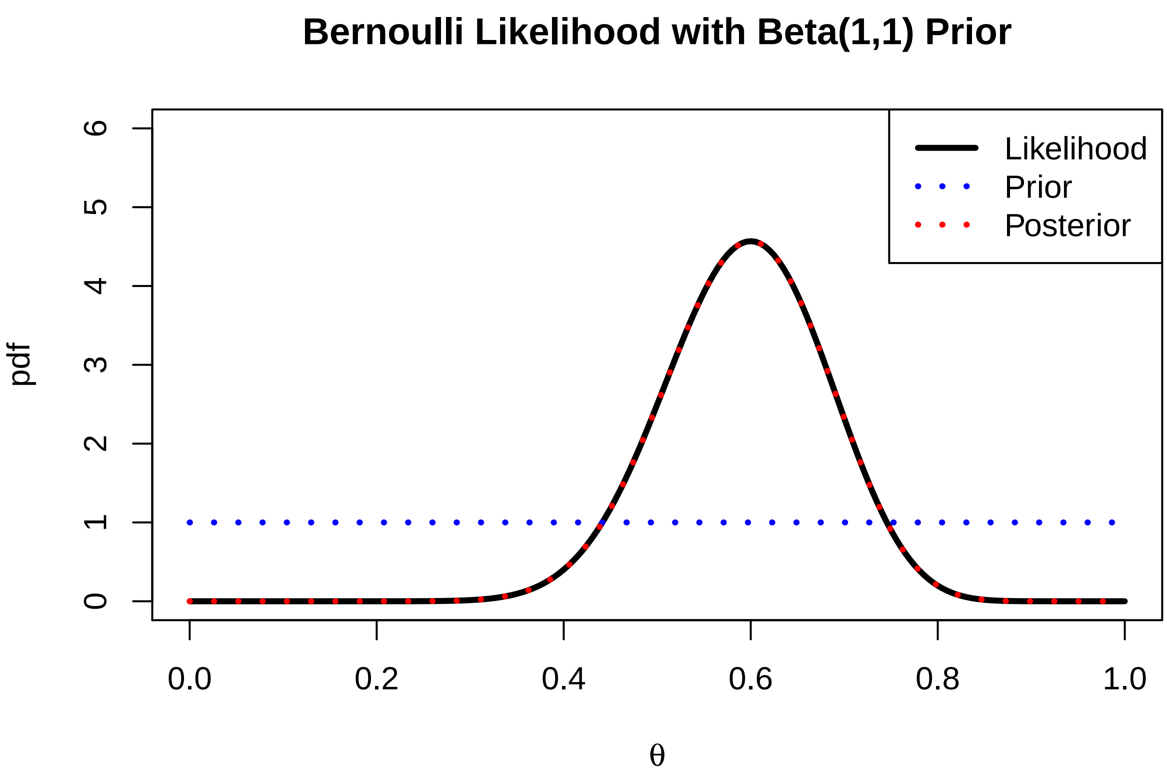
$$\theta \,|\, x_1, \ldots, x_n \sim \mathrm{Beta}(\alpha + k, \beta + n - k).$$

So, we know the normalizing constant is $C = \mathrm{B}(\alpha + k, \beta + n - k)$.

Here is an example in R. Let's say we observe 30 coin flips, 18 heads (the 1 values) and 12 tails (the 0 values). We will use a uniform, a.k.a., $\mathrm{Beta}(1, 1)$, prior for the probability of heads. Not surprisingly, the posterior is the same function as the likelihood in this case (after normalizing).

```r
## 18 ones and 12 zeroes
k = 18
n = 30

## x-axis for plotting
numSteps = 200
x = seq(0, 1, 1 / numSteps)

## Likelihood function
L = x^k * (1 - x)^(n - k)

## Just normalize likelihood to integrate to one (for purposes of plotting)
```

```
L = L / sum(L) * numSteps


### Uniform Prior
## Plot likelihood

plot(x, L, type = 'l', lwd = 3, ylim = c(0,6),
     main = "Bernoulli Likelihood with Beta(1,1) Prior",
     xlab = expression(theta), ylab = "pdf")

## Plot Beta(1,1) prior
lines(x, dbeta(x, 1, 1), lty = 3, lwd = 3, col = "blue")

## Plot posterior
lines(x, dbeta(x, k + 1, n - k + 1), lty = 3, lwd = 3, col = "red")

legend("topright", c("Likelihood", "Prior", "Posterior"),
       lty = c(1, 3, 3), lwd = 3, col = c("black", "blue", "red"))
```



Now, using the cdf of the Beta distribution, we can get a posterior probability interval for θ. This is sometimes called a *credible interval*. Let's take the "middle" 95% posterior interval. Then we want the interval from the 0.025 quantile to the 0.975 quantile of the posterior, $p(\theta \,|\, x_1, \ldots, x_n)$. In R this is computed like so:

```
qbeta(c(0.025, 0.975), k + 1, n - k + 1)
```

```
## [1] 0.4218696 0.7545240
```

The way we interpret this result is to say, "there is a 95% posterior probability that θ is in (0.42, 0.75), given the data we have observed." You might compare this Bayesian credible interval with the many different confidence intervals that have been proposed for a Bernoulli parameter:

https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

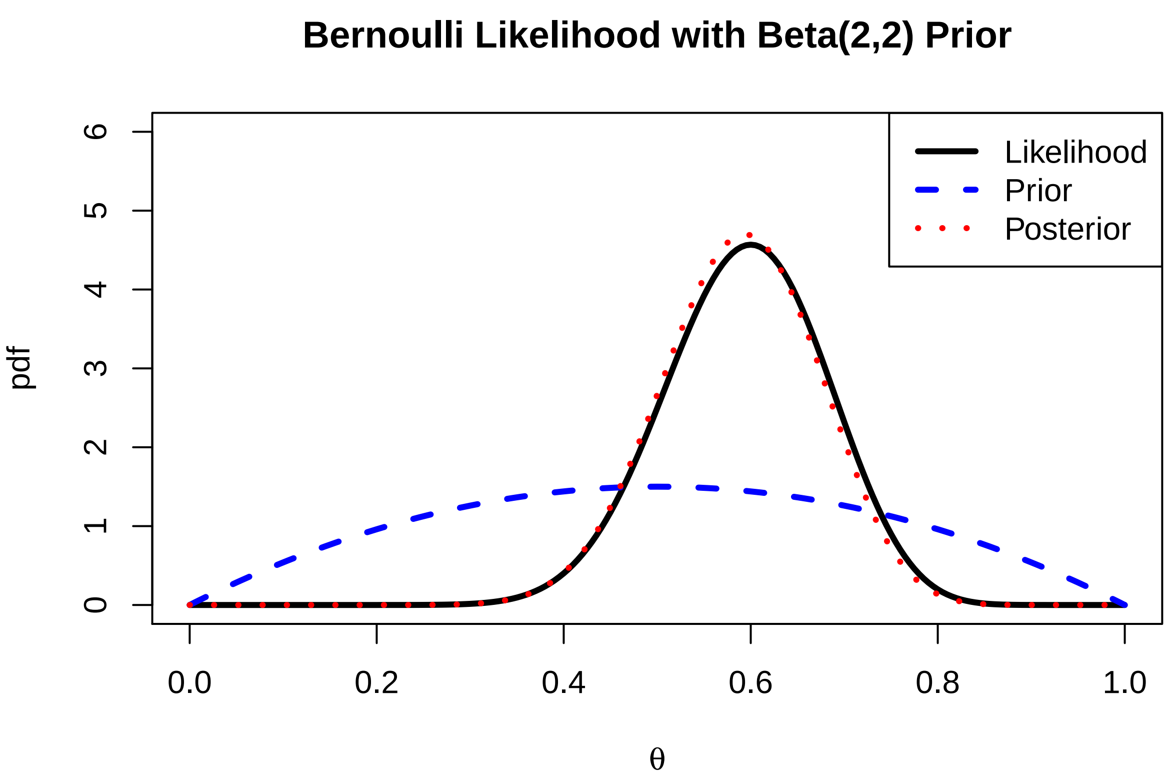
Notice we can answer other questions using the posterior probability, for example, what is the posterior probability that heads are more likely than tails, i.e., what is $P(\theta \geq 0.5 \,|\, x_1, \ldots, x_n)$? We compute this in R using the `pbeta` function, which is the cdf for the Beta distribution:

```
1 - pbeta(0.5, k + 1, n - k + 1)
```

```
## [1] 0.8594792
```

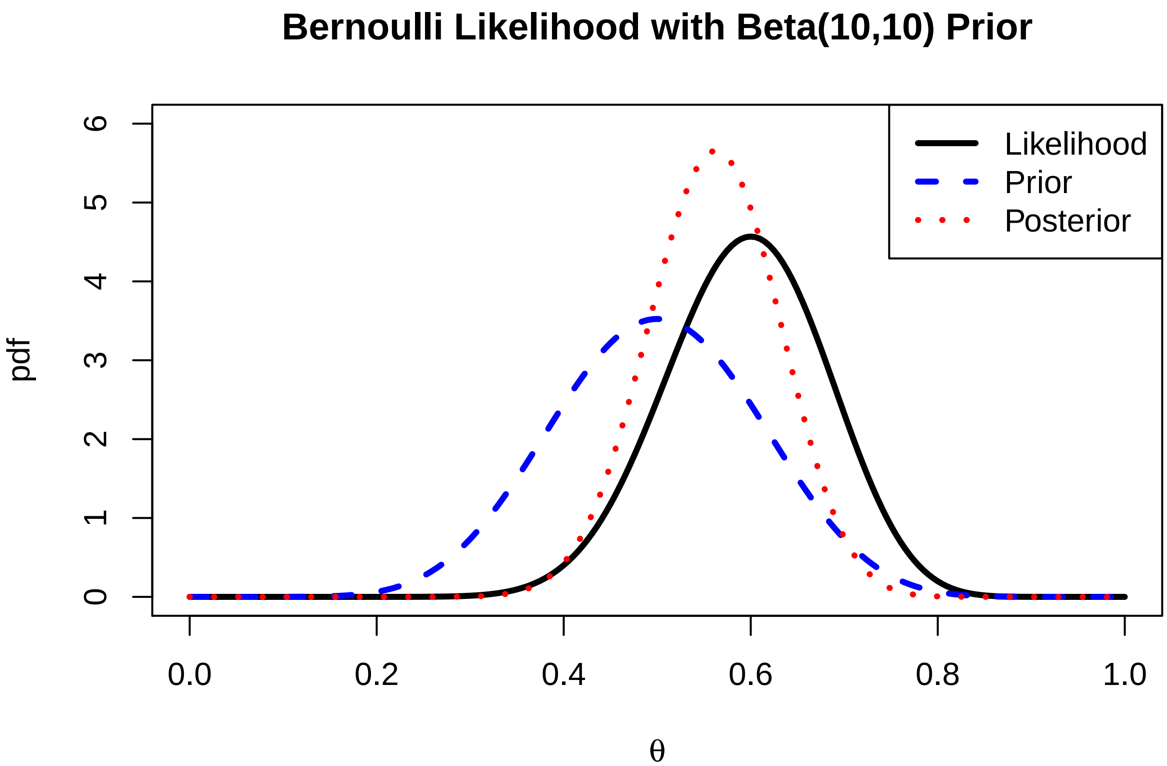
Next, let's see what happens if we use a different Beta prior. Now the shape of the prior is not uniform, and it has an effect on the posterior distribution. The posterior is a "compromise" between the likelihood function from the observed data and the prior. As we saw above, the Beta(α, β) prior can be thought of as adding additional observations to the likelihood (α "ones" and β "zeroes").

```
### Beta(2, 2) Prior
plot(x, L, type = 'l', lwd = 3, ylim = c(0,6),
     main = "Bernoulli Likelihood with Beta(2,2) Prior",
     xlab = expression(theta), ylab = "pdf")

## Plot Beta(2,2) prior
lines(x, dbeta(x, 2, 2), lty = 2, lwd = 3, col = "blue")

## Plot posterior
lines(x, dbeta(x, k + 2, n - k + 2), lty = 3, lwd = 3, col = "red")

legend("topright", c("Likelihood", "Prior", "Posterior"),
       lty = c(1, 2, 3), lwd = 3, col = c("black", "blue", "red"))
```



Increasing α, β will make the prior have more influence

```
### Beta(10, 10) Prior
plot(x, L, type = 'l', lwd = 3, ylim = c(0,6),
     main = "Bernoulli Likelihood with Beta(10,10) Prior",
     xlab = expression(theta), ylab = "pdf")

## Plot Beta(10,10) prior
lines(x, dbeta(x, 10, 10), lty = 2, lwd = 3, col = "blue")

## Plot posterior
lines(x, dbeta(x, k + 10, n - k + 10), lty = 3, lwd = 3, col = "red")

legend("topright", c("Likelihood", "Prior", "Posterior"),
       lty = c(1, 2, 3), lwd = 3, col = c("black", "blue", "red"))
```



And vice versa, increasing our sample size for the observed data will make the likelihood term more dominant.

```
### Beta(10, 10) Prior, but increase n by 10
n = 300
k = 180

L = x^k * (1 - x)^(n - k)
L = L / sum(L) * numSteps

plot(x, L, type = 'l', lwd = 3, ylim = c(0,15),
     main = "Bernoulli Likelihood with Beta(10,10) Prior (increased n)",
     xlab = expression(theta), ylab = "pdf")

## Plot Beta(10,10) prior
lines(x, dbeta(x, 10, 10), lty = 2, lwd = 3, col = "blue")
```

```r
## Plot posterior
lines(x, dbeta(x, k + 10, n - k + 10), lty = 3, lwd = 3, col = "red")

legend("topright", c("Likelihood", "Prior", "Posterior"),
       lty = c(1, 2, 3), lwd = 3, col = c("black", "blue", "red"))
```