# Data 582 - Bayesian Inference

Lab 3: Beta-Binomial Model

## Contents



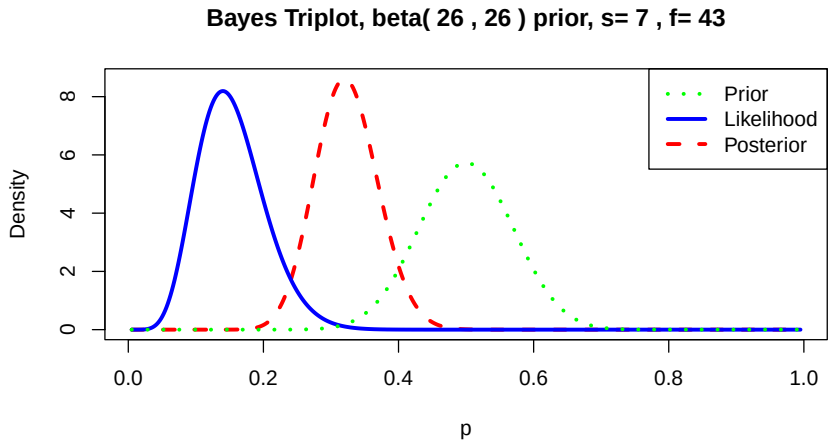
## 1 Introduction

In lecture we saw how we could perform Bayesian Inference on a Binomial proportion. Here we will look at some examples involving this topic. Recall that the beta distribution is a conjugate prior for the corresponding $\text{Bin}(n, \theta)$ data model, i.e. the resulting posterior is also a beta distribution. More specifically, if we adopt a beta prior $\theta \sim \text{Beta}(a, b)$ and we observe $y$ successes in $n$ Bernoulli trials, then our posterior has the form of $p(\theta \mid y) \sim \text{Beta}(a + y, b + n - y)$. For instance if we let:

$$\begin{aligned} p(\theta) &\sim \text{Beta}(26, 26) \\ p(y \mid \theta) &\sim \text{Beta}(7 + 1, 43 + 1) \\ p(\theta \mid y) &\sim \text{Beta}(33, 69) \end{aligned}$$

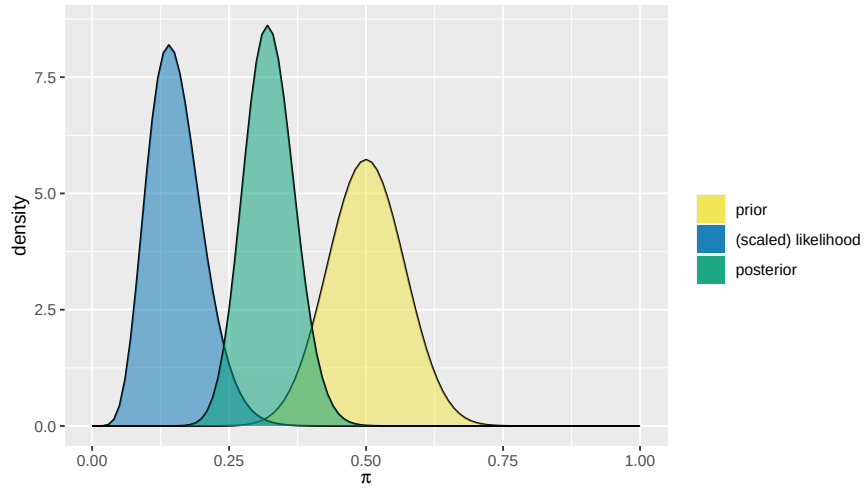
To see the tug-of-war between the prior and likelihood to arrive at the posterior, we can plot them on a single graph using, for example, the `triplot` function from the `LearnBayes` package.

```
library(LearnBayes)
a = 26; b=26   # prior parameter values
y = 7; n=50    # number of success; number of Bernoulli trials
triplot(prior=c(a, b), data=c(y,n-y))
```




The above is just one example of a package that can perform this Beta-Binomial modelr. Another options includes the **bayesrule** package:

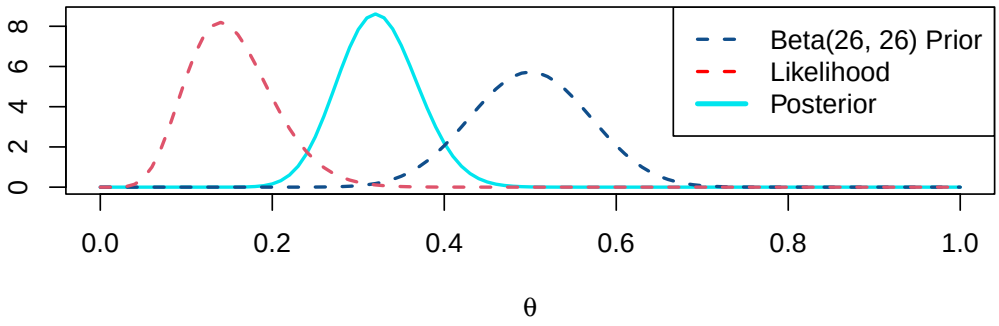
```
library(bayesrules)
plot_beta_binomial(alpha = 26, beta = 26, y = 7, n = 50)
```




Alternatively, we could produce this “from scratch" using the `dbeta` functions. While it is more tedious, it will be beneficial to code at least once to get a deeper understanding of whats going on the “under-the-hood".

```
# plot the posterior:
curve(dbeta(x, shape1 = a+y, shape2 = b+n-y), xlab=expression(theta),
      lwd=2, col="turquoise2", ylab="", xlim=c(0,1))
# plot the prior:
curve(dbeta(x, a, b), lwd=2, col="dodgerblue4",lty=2, add=TRUE)
# plot the likelihood:
curve(dbeta(x, y+1, n-y+1), lwd= 2, col=2, lty =2 , add=TRUE)

prior.name <- paste("Beta(", a, ", ", b,") Prior",sep="")
```

```
legend("topright", legend = c(prior.name, "Likelihood", "Posterior"),
       lwd=c(2,2,3), lty=c(2,2,1), col=c("dodgerblue4", "red", "turquoise2"))
```




Notice above how we are using the `dbeta` function for plotting the likelihood. Recall from lecture that the likelihood for this data is given by:

$$\mathcal{L}(\theta) \propto \theta^y (1-\theta)^{n-y}$$

which has the same functional form as Beta distribution with shape parameters $\alpha = y+1$ and $\beta = n-y+1$. For that reason, you may see this likelihood being referred to as a "Beta likelihood". Another way to put this is $y \mid \theta \sim \text{Beta}(y+1, n-y+1)$, that is, $y$ given $\theta$ follows the Beta distribution with shape parameters $\alpha = y+1$ and $\beta = n-y+1$.

# 2 Point Estimates

While graphical representation of the posterior contains all the current information about the unknown parameter, numeric summaries are useful and needed as well.

## 2.1 The Frequentist Point Estimate (MLE)

While we saw an example of finding the MLE for a binomial proportion for a particular example involving coin flips in lecture 2, more generally we can derive the frequestist point estimate for $\theta \ldots$

$$\begin{aligned} \mathcal{L}(\theta) &= \theta^y (1-\theta)^{n-y} \\ \implies \log \mathcal{L}(\theta) = \ell(\theta) &= y \log \theta + (n-y) \log 1 - \theta \end{aligned}$$

taking the derivative w.r.t. $\theta$, setting =0 and solving for $\theta$ gives us the Frequents MLE

$$\begin{aligned}
\frac{y}{\theta} + (-1)\frac{n-y}{1-\theta} &= 0 \\
\frac{y}{\theta} &= \frac{n-y}{1-\theta} \\
y - y\theta &= n\theta - y\theta \\
\implies \theta &= \frac{y}{n}
\end{aligned}$$

i.e. the MLE = $\theta_{MLE} = \frac{y}{n}$, is the value of the $\theta$ that maximizes the likelihood function.

## 2.2 The Bayesian Point Estimate

To obtain a point estimate for $\theta$ we could use:

- posterior maximum
- posterior median
- posterior mean (typically our point estimate of choice)

### The Posterior Mean

One popular method of describing Bayesian point estimate for a parameter $\theta$ is the take the mean of the posterior distribution $p(\theta \mid y)$. A quick Wikipedia search tells us that the mean of a Beta($\alpha, \beta$) distribution is

$$\mu = \frac{\alpha}{\alpha + \beta}$$

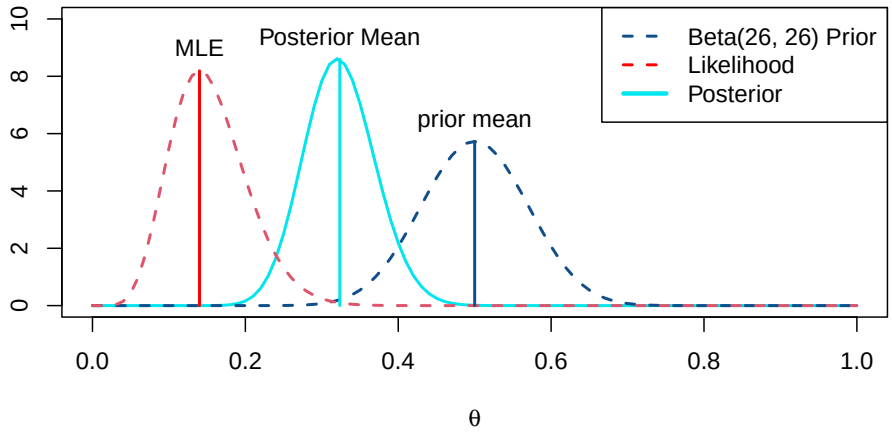
For a beta prior, $\theta \sim \text{Beta}(a, b)$, and binomial likelihood, $y \sim \text{Beta}(y+1, n-y+1)$, which has the posterior distribution $\theta \mid y \sim \text{Beta}(a+y, b+n-y)$, the posterior mean for the Binomial proportion $\theta$ is

$$E[\theta \mid y] = \mu = \frac{a+y}{a+y+b+n-y} = \frac{a+y}{a+b+n} \tag{1}$$

In our example, with the Beta$(26, 26)$ prior,

$$E[\theta \mid y] = \frac{a+y}{a+b+n} = \frac{26+7}{26+26+50} = \frac{33}{102} = 0.3235294$$

For a beta prior and binomial likelihood, the posterior mean is always between the prior mean, $\frac{a}{a+b}$, and the MLE, $\frac{y}{n}$.




While we can see this in the visualization above, we can also see mathematically how the posterior mean is a weighted average of the prior mean and Frequentist MLE (i.e sample success rate):

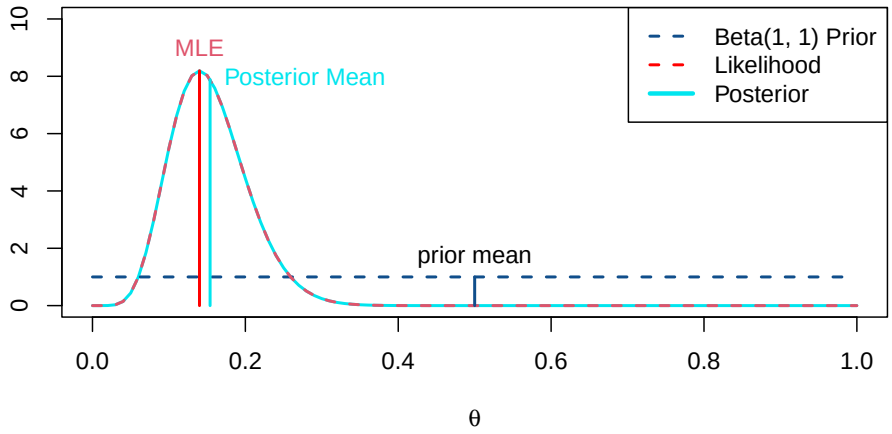
$$\begin{aligned} E(\theta \mid Y = y) &= \frac{a}{a+b+n} + \frac{y}{a+b+n} \\ &= \frac{a}{a+b+n} \cdot \frac{a+b}{a+b} + \frac{y}{a+b+n} \cdot \frac{n}{n} \\ &= \frac{a+b}{a+b+n} \cdot \frac{a}{a+b} + \frac{n}{a+b+n} \cdot \frac{y}{n} \\ &= \frac{a+b}{a+b+n} \cdot E(\theta) + \frac{n}{a+b+n} \cdot \frac{y}{n}. \\ &= c \times \text{Prior mean} + (1-c) \times \text{MLE}. \end{aligned}$$

where $c = \frac{a+b}{a+b+n}$ and $(1-c) = \frac{a+b+n}{a+b+n} - \frac{a+b}{a+b+n} = \frac{n}{a+b+n}$

Notice how when we use a Uniform prior, i.e. Beta(1,1) prior, the posterior is simply proportional to the likelihood:

$$\begin{aligned} p(\theta \mid y) &\propto p(\theta) p(y \mid \theta) \\ &\propto p(\theta) \mathcal{L}(\theta) \\ &\propto \theta^{1-1}(1-\theta)^{1-1} \times \theta^{7}(1-\theta)^{50-7} \\ &\propto \theta^{7}(1-\theta)^{43}, \qquad \text{for } 0 \leq \theta \leq 1 \end{aligned}$$

While graphically, the posterior and likelihood are identical, note that the MLE = $\frac{y}{n} = \frac{7}{50} = 0.14$ is *not* equal to the Bayesian point estimate $\frac{a+y}{a+b+n} = \frac{1+7}{1+1+50} = 0.1538462$.




**Other Bayesian Point Estimates**

The posterior median and posterior mode are sometimes used instead of the posterior mean as Bayesian point estimates. For example, the mode of Beta($\alpha$, $\beta$) distribution is given by:

$$\tilde{\mu} = \frac{\alpha - 1}{\alpha + \beta - 2} \tag{2}$$

Notice that when we use a uniform prior on Binomial samples the posterior distribution of Beta(8, 44) yields a posterior mode point estimate of $\frac{7}{50} = 0.14$ which *is* equivalent to the MLE. Recall that the mode of a set of data values is the value that appears most often. Since the mode of a continuous probability distribution is often considered to be the value(s) for which its probability density function has a locally maximum value, i.e. any peak is a mode, then this result should not come as a surprise.

## 2.3 Bayesian Posterior Intervals

Bayesian *credible intervals* (sometimes called *creditable sets*) are analogous to the Frequentist confidence intervals.

**Bayesian Credible Interval**

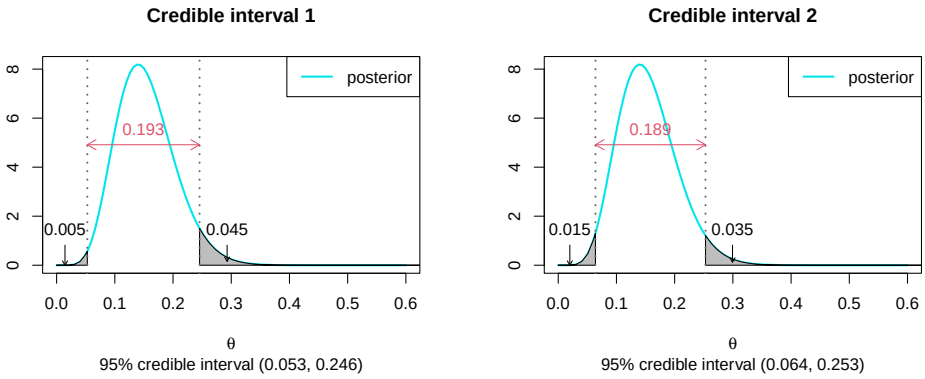
A $100(1-c)\%$ **Bayesian credible interval** (or simply *credible interval*) for $\theta$ is an interval $[a, b]$ having

$$P(a \leq \theta \leq b \mid Y = y) = 1 - c$$

Clearly, there is more than a single credible interval that satisfies the above condition. Let's look at our posterior $\theta \mid y \sim$Beta(8, 44). All of the following intervals correspond to a 95% credible interval; however, with differing widths (where width is taken to be the

upper bound minus the lower bound). A visualization comparing two of these intervals are given below.

| Lower | Upper | Width |
|---|---|---|
| 0.053 | 0.246 | 0.193 |
| 0.064 | 0.253 | 0.189 |
| 0.070 | 0.263 | 0.192 |
| 0.075 | 0.276 | 0.201 |
| 0.079 | 0.304 | 0.225 |




To verify that the above credible intervals captures 95% posterior probability, we can use the `pbeta` function.. For example, we can check that the first credible interval contains 95% posterior probability by:

```
pbeta(0.246, 8, 44) - pbeta(0.053, 8, 44)

## [1] 0.9504818
```

Note these are not exactly 95% due to rounding error.

There are two commonly used ways to produce numeric posterior summaries: Central Intervals (i.e. equal-tail posterior credible sets), or Highest Posterior Density Regions.

### Highest Posterior Density Regions

Highest Posterior Density Regions is credible interval (for some given confidence level) having the shortest width.

**Highest posterior density (HPD) region**

The $100(1-c)\%$ **highest posterior density (HPD) region** for $\theta$ is the narrowest interval $[a, b]$ having

$$P(a \leq \theta \leq b \mid Y = y) = 1 - c$$

While these intervals are desirable (esp. for skewed distributions), they are harder to find in practice.

### Central Intervals

A *central interval* (sometimes called an *equal-tailed interval*) is one in which the probability of being below the interval is as likely as being above it. For example, the endpoints of a 95% equal-tail interval are the 0.025 and the 0.975 quantiles of the posterior distribution.

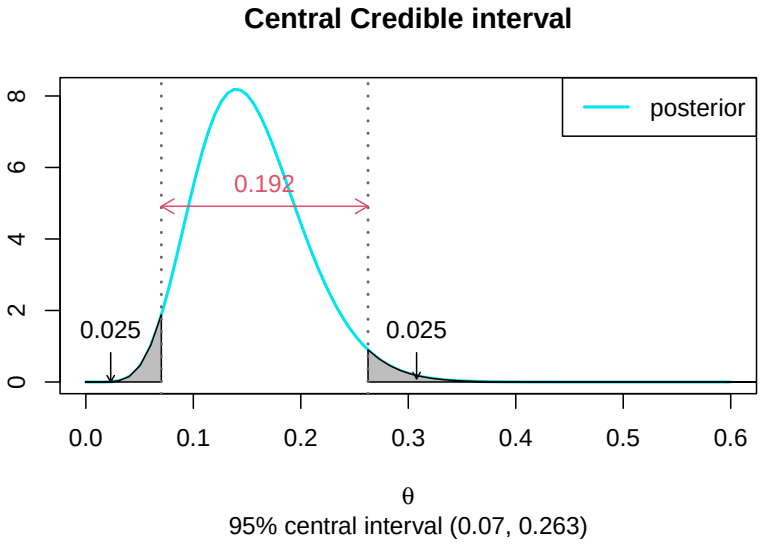
**Central/Symmetric Credible Interval**

A **central/symmetric** $100(1-c)\%$ **credible interval** for $\theta$ is an interval $[a, b]$ having

$$P(\theta \leq a \mid Y = y) = c/2$$
$$P(\theta \geq b \mid Y = y) = c/2$$

The values $a$ and $b$ above are the $c/2$ and $1 - c/2$ posterior quantiles of $\theta$.

Central intervals are easy to compute and easy to understand and the most commonly reported Bayesian posterior intervals.




We can use built-in R functions to calculate them. With a Beta$(a, b)$ prior, and posterior density Beta$(a + y, b + n - y)$, those quantiles are calculated in R using

```
# returns: c(0.025 quantile, 0.975 quantile)
qbeta(c(0.025, 0.975), a+y, b+n-y)

## [1] 0.07024083 0.26255154
```

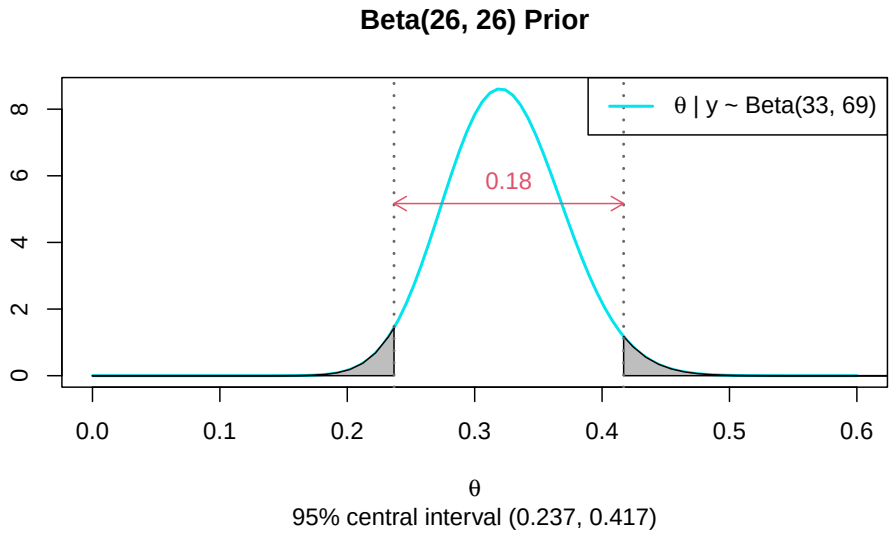
Below we display this interval for our Beta(26, 26) prior with Beta(33, 69) posterior. Note that the shaded gray region on the left and right both have the probability of 0.025.

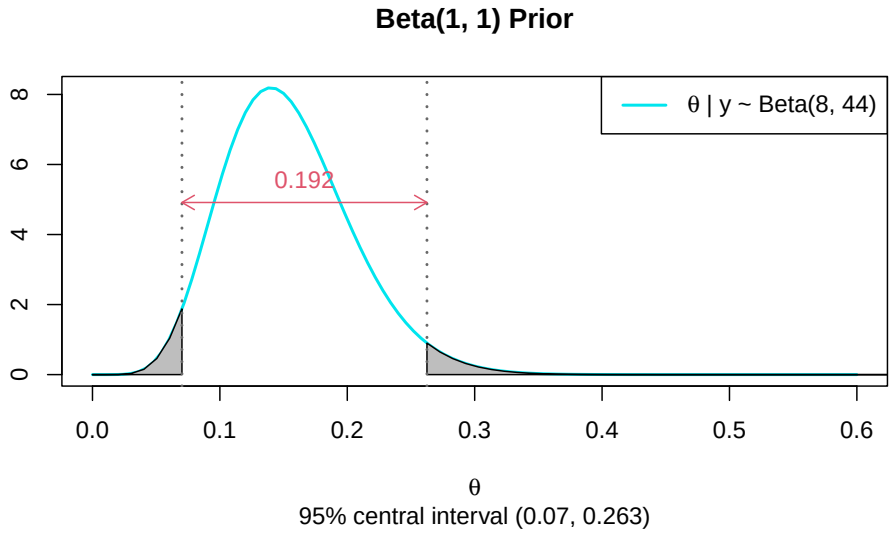
```
a = 26; b = 26 # hyperparameters
y = 7; n = 50 # data
aa = a +y ; bb = b + n -y # posterior parameters
```

```r
qbeta(c(0.025, 0.975), aa, bb)

## [1] 0.2366944 0.4169196

diff(qbeta(c(0.025, 0.975), aa, bb)) # width of central interval

## [1] 0.1802252
```

Notice how the 95% central interval for the uniform prior is wider

```r
a = 1; b = 1; aa = a +y ; bb = b + n -y
qbeta(c(0.025, 0.975), aa, bb)

## [1] 0.07024083 0.26255154

diff(qbeta(c(0.025, 0.975), aa, bb)) # width of central interval

## [1] 0.1923107
```






# 3 Examples

## 3.1 Happiness Data

**Example 3.1** (Happiness data). *Each female of age 65 or over in the 1998 General Social Survey was asked whether or not they were generally happy. Let $Y_i = 1$ if respondent $i$ reported being generally happy, and let $Y_i = 0$ otherwise. Suppose 129 individuals surveyed and 118 individuals report being generally happy (91%); 11 individuals do not report being generally happy (9%).*

Source: Section 3.1 Hoff (2009)

Let's perform Bayesian inference on $\theta$, the proportion of happy female senior citizens within this population. Let $y = \sum_{i=1}^{n} y_i$ be the number of individuals report being generally happy in $n$ female seniors surveyed. We will assume that $Y \sim$Binomial($n$,$\theta$), i.e. that the answers to this question are independent from one another and that the probability of being happy is equal to $\theta$ for all cases. Assuming a Beta prior on $\theta$ with shape parameters $\alpha, \beta > 0$ we have seen in class that the posterior distribution for $\theta$ works out to be

$$\theta \mid y \sim \text{Beta}(\alpha + y, \beta + n - y).$$

In R, we can use `dbeta()`, `pbeta()`, `qbeta()` and `rbeta()` to obtain the respective density function, cumulative distribution function (cdf), quantile function, and random observations from a beta distribution.

First lets set up the parameters of our model assuming a uniform prior on $\theta$ (that is a Beta prior with $\alpha = \beta = 1$):

```
n = 129    # number of bernoulli trials
y = 118    # number of successes in n bernoulli trials
a = 1      # shape 1 for the Beta prior on theta
```

```
b = 1      # shape 2 for the Beta prior on theta
```

**The likelihood**

For Binomial data, we know that the likelihood is given by:

$$p(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \quad y = 0, 1, 2, \ldots, n$$

The likelihood (up to a constant of proportionality) for this data is given by:

$$\mathcal{L}(\theta) \propto \theta^{118}(1-\theta)^{129-118} = \theta^{118}(1-\theta)^{11} \tag{3}$$

To plot the likelihood on the same scale as the Prior and the Posterior, we can normalize the above such that it integrates to 1. We can arrive at this using calculus or recognize that (3) has the functional form of a Beta distribution with $\alpha$= 119 and $\beta$ = 12. Hence our normalizing constant should be equal to $1/\mathrm{B}(\alpha, \beta)$ where

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \frac{\Gamma(119)\Gamma(12)}{\Gamma(119+12)} = \frac{\Gamma(119)\Gamma(12)}{\Gamma(131)}$$

This can be calculated using the `beta()` or `gamma()` function in R:

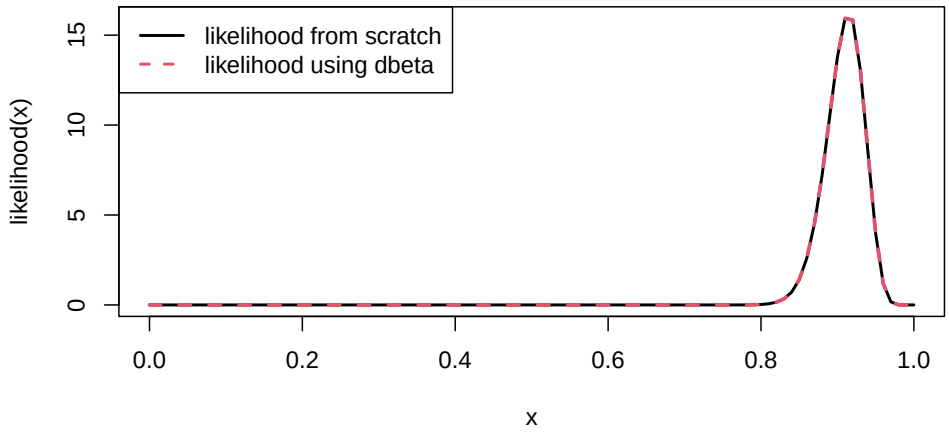
```
# normalizing constant for the likelihood
# = 1/B(a,b) = 1/(gamma(a)*gamma(b)/gamma(a+b))
# = gamma(a+b)/(gamma(a)*gamma(b))
alpha <- a +y
beta <- b+n-y
1/beta(alpha, beta) # same as below

## [1] 3.458373e+17

(likeC <- gamma(alpha+beta)/(gamma(alpha)*gamma(beta)))

## [1] 3.458373e+17
```

We can plot the normalized likelihood “from scratch" using:

```
likelihood <- function(theta) likeC*theta^y*(1-theta)^{n-y}
curve(likelihood(x), lwd = 2)
```

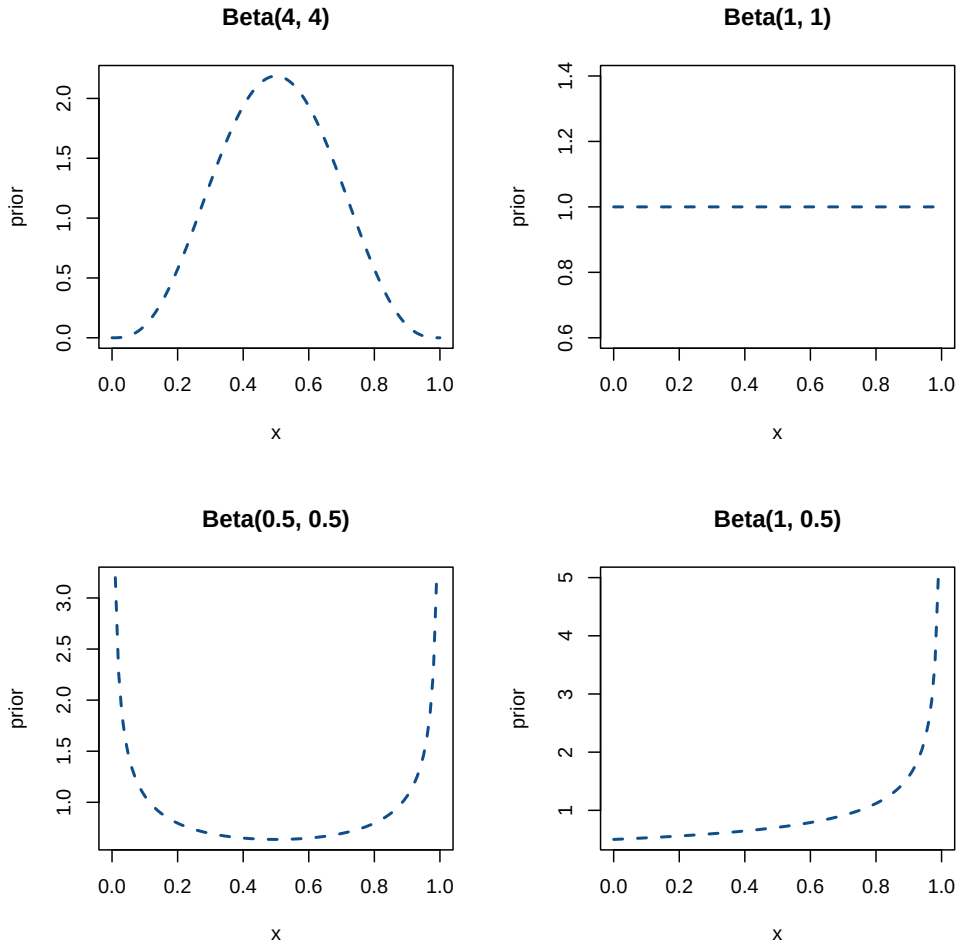
or, more simply, we could use the `dbeta` function:

```r
curve(dbeta(x, shape1 = y+1, shape2 = n-y+1), add = T, col = 2, lwd = 2.5, lty = 2)
```




**The prior**

To plot the prior, again we can use the `dbeta()` function. The parameters are going to dictate the shape of the prior distribution. Let's investigate a few examples:

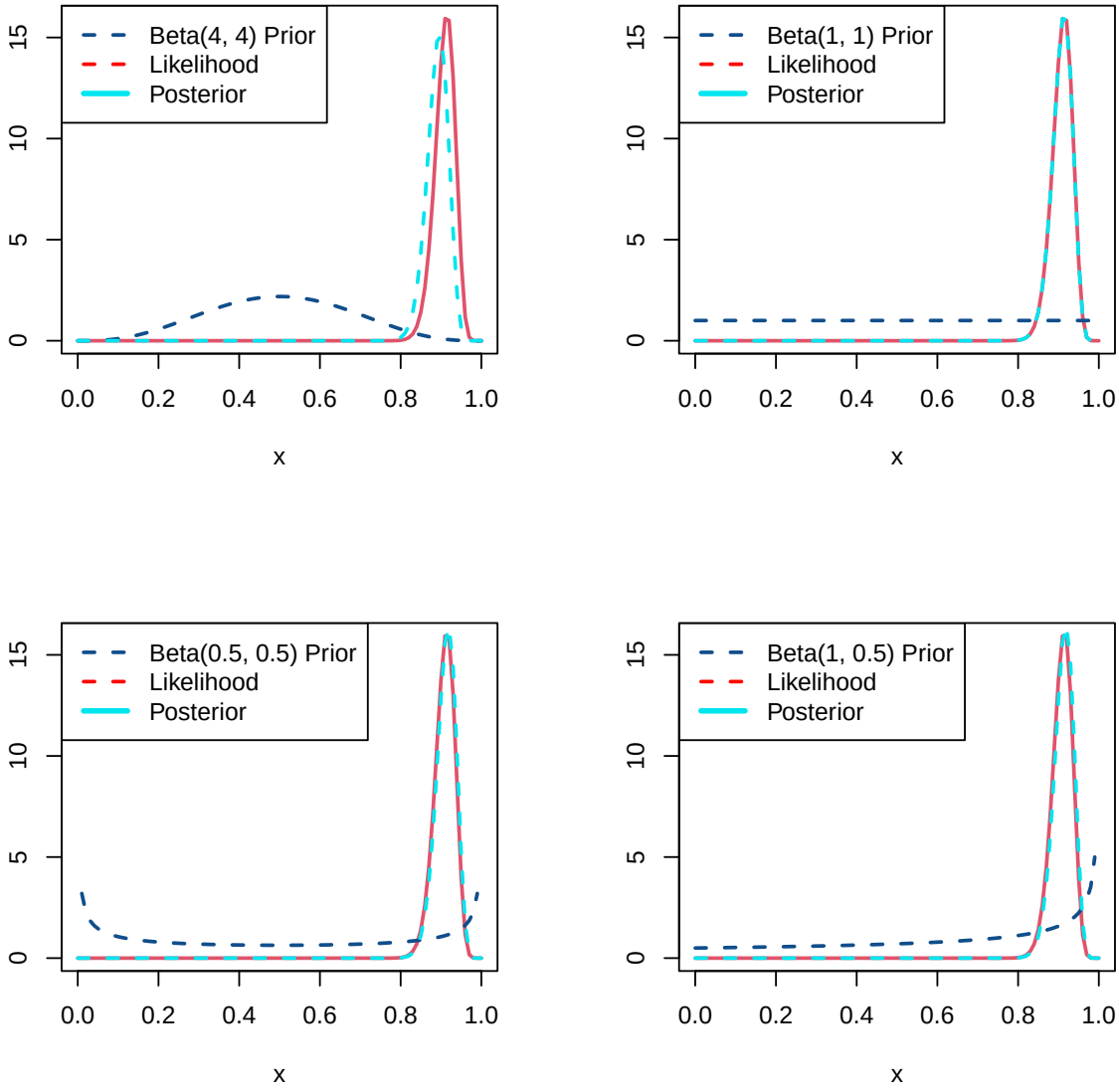



**The posterior**

In class we have seen how the beta family of priors is conjugate for the binomial likelihood, since the posterior distribution is also beta—a member of the same parametric family as the prior distribution. More specifically, the posterior distribution was found (up to a constant of proportionality) to be:

$$p(\theta \mid y) \propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}$$

Following the same logic from the likelihood, we can by-pass the caclulations for finding the normalizing constant by recognizing that this has the function form of a Beta distribution. Hence $\theta \mid y \sim \text{Beta}(\alpha + y, \beta + n - y)$. Again we can use the `dbeta()` function to plot this distribution. Summarizing this all into one graphic we get:




## 3.2 Bayesball

Suppose $X_1, \ldots, X_n$ represent a hit (1) or strikeout (0) during "at bats" by a made-up MLB rookie Jacob (count walks as a hit). We will assume independence for "at bats", and we can consider each $X_i$ as a Bernoulli trial with probability of success $p$. We could therefore interpret $p$ as Jacob's "true" batting average.

We know that $0 \leq p \leq 1$ where if $p = 0$ then Jacob will never get a hit and if $p = 1$ then Jacob will always gets a hit. Note that an "average" hitter tends to have a batting average around 0.25 for their career (1 in 4 at bats result in a hit), and a top tier hitter can end up around 0.33 (1 in 3 at bats).

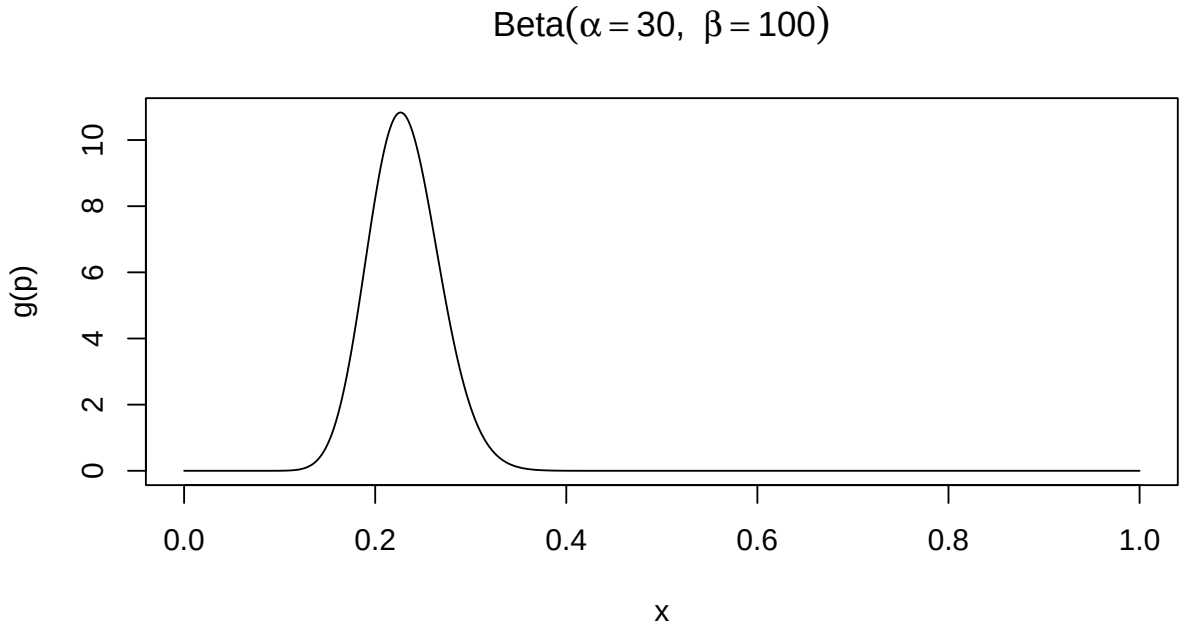
We will walk through this thought experiment in both the Bayesian and frequentist

paradigms while contrasting results.

**Strong priors**

Under the Bayesian paradigm, we assume $p$ itself is a random variable. For this particular example, my write-up suggests that we have pretty solid information about what $p$ 'should' look like — this is an ideal scenario for the Bayesian approach.

To summarize, we know that $p$ is bounded by [0, 1] and is much, much more likely to fall somewhere in the lower third of that interval (but probably not in [0, .1] for an MLB player). I'll suggest that $p \sim \text{Beta}(\alpha = 30, \beta = 100)$ is a sensible prior. This assumes, before Jacob ever steps up to the plate, that his true batting average is described by this probability density:




**Jacob's first game**

Jacob has 3 hits during 5 at bats in his first MLB game.

The frequentist paradigm views $p$ as a fixed, but unknown, value. Our previous work suggests that

$$\hat{p}_f = \frac{\sum x_i}{n}$$

is a good estimator for this paradigm. So our frequentist point estimate for Jacob is

$$\hat{p}_f = 3/5 = 0.6$$

As discussed in class, since the beta distribution is a conjugate prior we know that

the Bayesian estimate can be given by

$$\hat{p}_b = \frac{\sum x_i + \alpha}{n + \alpha + \beta}.$$

So our Bayesian point estimate for Jacob is

$$\hat{p}_b = \frac{3 + 30}{5 + 30 + 100} = 0.244$$

Which approach is better?

It is easy to argue that the Bayesian approach is more likely to be closer to Jacob's career batting average in the long run (since no player in history has come even close to batting at 0.6 for their career). The Bayesian estimator is therefore incorporating our prior belief that Jacob is extremely unlikely to reach that mark in the longrun.

Furthermore, we have only seen Jacob play one game! This is a small sample size to estimate from. The Bayesian estimator is therefore 'forcing' our estimate of Jacob's batting average closer to that of an average player through our prior distribution on $p$.

**Jacob's first season**

So what if Jacob has a phenomenal first season? Let's say 238 hits during 684 at bats. This would have made him the best in the league for the 2018 season.

Frequentist:

$$\hat{p}_f = \frac{238}{684} = 0.348$$

Bayesian:

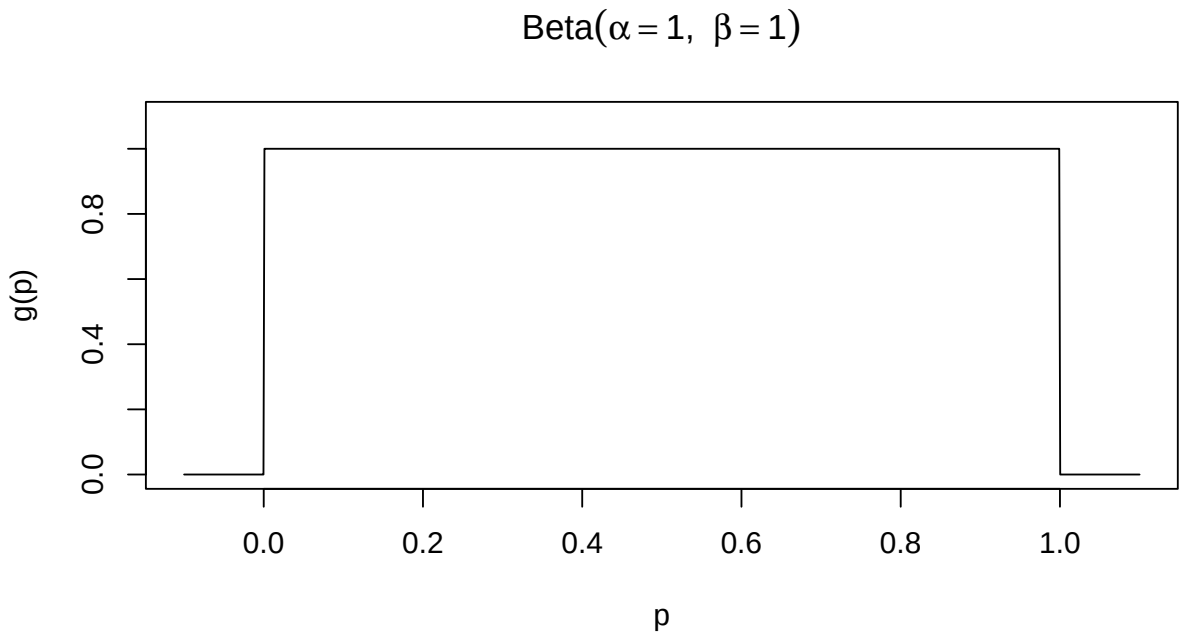
$$\hat{p}_b = \frac{238 + 30}{684 + 30 + 100} = 0.329$$

As you can see, while the prior distribution is still pulling Jacob's average towards that of an average player, the larger sample of observed data is much more heavily influencing the estimate.

One can see as we move towards larger and larger sample sizes, the frequentist and Bayesian paradigms converge towards agreement on the point estimate.

**Uninformative priors**

What if we pretend that we are ignorant about the general form of batting averages, aside from being bounded on [0,1]? Suppose we assume $p \sim \text{Beta}(\alpha = 1, \beta = 1)$, which is equivalent to a standard uniform, ie. $Unif(0, 1)$:




Let's say Jacob goes 1 for 5 in his first game:
Frequentist:

$$\hat{p}_f = \frac{1}{5} = 0.2$$

Bayesian:

$$\hat{p}_b = \frac{1+1}{5+1+1} = 0.286$$

Which are fairly close even on a small sample size, though its worth noting that the uniform prior is pulling the Bayes estimator somewhat towards the expected value of that prior (in this case, the value 0.5).
**Note:** While these are often called uninformative, or ignorant, priors, some statisticians argue against that wording, and even their usage. The crux of the argument is that a uniform prior distribution actually assumes quite a bit about your parameter. Namely, that each value of the parameter is equally likely.