Stat 343: Posterior Exploration

Introduction

Let's estimate the proportion of M&M's that are blue. Call this proportion θ. Suppose we take a sample of n M&M's and let the random variable X denote a count of how many are blue in that sample. Our model is $X \overset{\text{iid}}{\sim} \text{Binomial}(n, \theta)$.

We have developed two approaches to inference for θ:

1. The maximum likelihood estimate $\hat{\theta}_{MLE} = \frac{x}{n}$, which maximizes the likelihood function $\mathcal{L}(\theta|x)$.
2. A Bayesian approach with conjugate prior distribution given by $\Theta \sim Beta(a, b)$. The posterior distribution is given by $\Theta|n, x \sim Beta(a + x, b + n - x)$. From this posterior distribution, we can obtain point estimates (such as the posterior mean, posterior median, posterior mode) and interval estimates (such as a posterior 90% credible interval).

We have three related goals in this lab:

1. To see what the posterior distribution looks like in Bayesian inference, and how this changes as the sample size n increases.
2. To see what effect the prior distribution has on Bayesian inferences, and how this changes as the sample size n increases.
3. To compare maximum likelihood and Bayesian estimates of θ, and see how these estimates change as the sample size n increases.

Procedure

Step 0. Prior Specifications

In problem set 2, you found values of the parameters a and b for a beta distribution that represented your prior beliefs about the proportion of M&M's that are blue, before looking at any data.

In order to get answers that are the same for everyone, and to understand the relationship between the prior distribution and the posterior, let's all work with the following three specifications of prior distributions:

- Non-informative Prior: a = 1 and b = 1
- Informative Prior, weak prior knowledge: a = 2, b = 8
- Informative Prior, strong prior knowledge: a = 20, b = 80

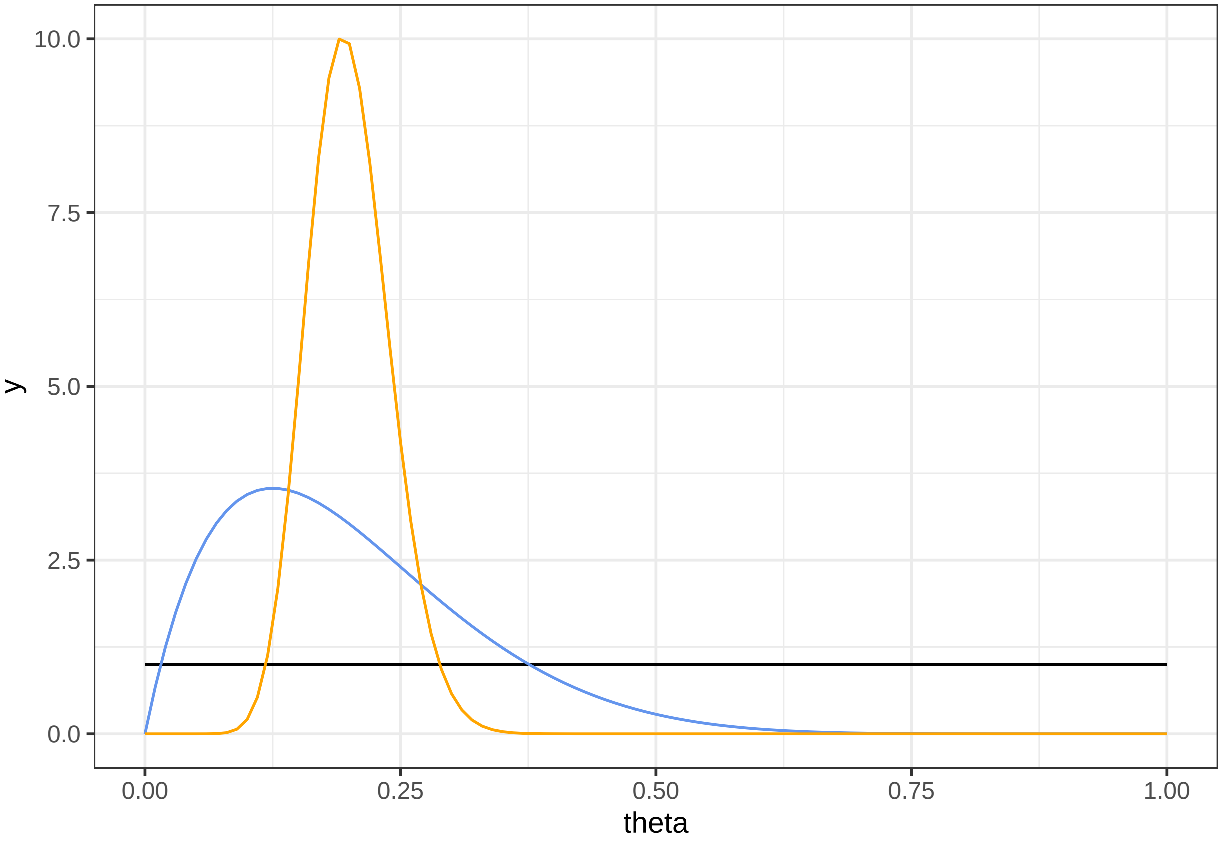
The pdfs for these three prior distributions are below:

```r
a_noninformative <- 1
b_noninformative <- 1

a_weakly_informative <- 2
b_weakly_informative <- 8

a_strongly_informative <- 20
b_strongly_informative <- 80

ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_noninformative, shape2 = b_noninformative)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_weakly_informative, shape2 = b_weakly_informative),
    color = "cornflowerblue") +
  stat_function(fun = dbeta,
    args = list(shape1 = a_strongly_informative, shape2 = b_strongly_informative),
    color = "orange") +
  theme_bw()
```



The prior means and a 95% prior credible interval based on the non-informative prior are calculated below:

```
a_noninformative/(a_noninformative + b_noninformative)
```

```
## [1] 0.5
```

```
qbeta(c(0.025, 0.975), shape1 = a_noninformative, shape2 = b_noninformative)
```

```
## [1] 0.025 0.975
```

The prior means and a 95% prior credible interval based on the weakly informative prior are calculated below:

```
a_weakly_informative/(a_weakly_informative + b_weakly_informative)
```

```
## [1] 0.2
```

```
qbeta(c(0.025, 0.975), shape1 = a_weakly_informative, shape2 = b_weakly_informative)
```

```
## [1] 0.02814497 0.48249651
```

The prior means and a 95% prior credible interval based on the strongly informative prior are calculated below:

```
a_strongly_informative/(a_strongly_informative + b_strongly_informative)
```

```
## [1] 0.2
```

```
qbeta(c(0.025, 0.975), shape1 = a_strongly_informative, shape2 = b_strongly_informative)
```

```
## [1] 0.1279847 0.2833676
```

Make sure you understand what the different prior distributions say about the analyst's beliefs about the value of θ before moving on.

Step 1. Take some samples and record data

In order to compare the estimates above, let's take samples of a few different sizes and plot the likelihood function, the posterior distribution based on two different prior distributions, and point estimates from each method for each sample size.

Take about 50 M&M's – the exact number is not important.

Record the following:

Was your first M&M blue? (Pick a random M&M from your sample that you will count as your first draw)

no; X=0

Out of your first 10 M&M's, how many were blue?

1

Out of your first 20 M&M's, how many were blue?

2

Out of all of the M&M's you picked, how many were blue? Also, how big was your total sample size?

x = 10 n = 46

Step 2. Find posterior distribution, point and interval estimates

Sample of size n = 1

The code below is exactly the same as the code above for exploring the prior distributions. Update the code to explore the posterior distributions obtained from each prior specification based on the data observed for your sample of size 1, and compare to the maximum likelihood estimate.

You will have to make two changes:

1. Update the a and b parameters to be the posterior parameters corresponding to each prior.
2. Add in a vertical line at the maximum likelihood estimate (use `geom_vline(xintercept = *)`, replacing the * with the maximum likelihood estimate).

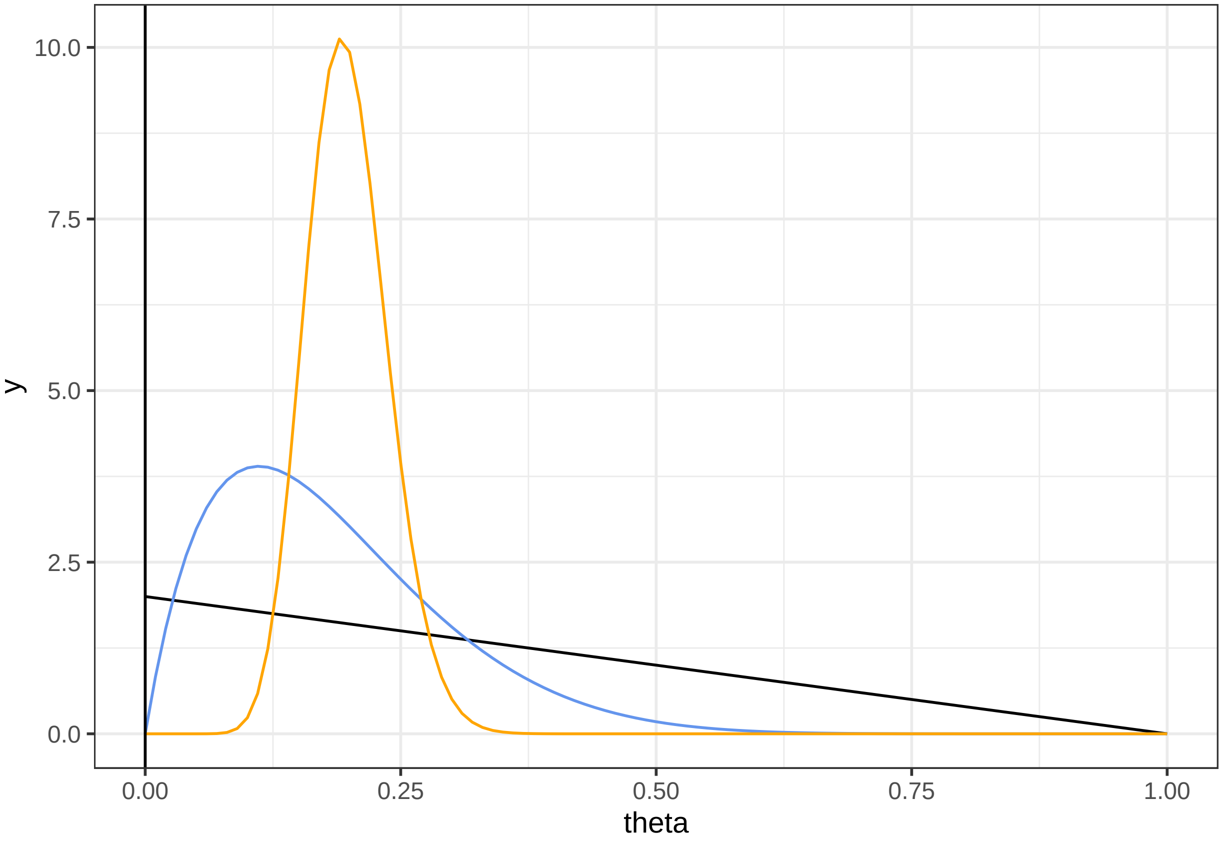
The rest of the code can be left as is.

```
a_noninformative <- 1 + 0
b_noninformative <- 1 + (1 - 0)

a_weakly_informative <- 2 + 0
b_weakly_informative <- 8 + (1 - 0)

a_strongly_informative <- 20 + 0
b_strongly_informative <- 80 + (1 - 0)
```

```r
# I observed X = 0 for my sample of size 1
ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_noninformative, shape2 = b_noninformative)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_weakly_informative, shape2 = b_weakly_informative),
    color = "cornflowerblue") +
  stat_function(fun = dbeta,
    args = list(shape1 = a_strongly_informative, shape2 = b_strongly_informative),
    color = "orange") +
  geom_vline(xintercept = 0) +
  theme_bw()
```



Prior mean and 95% posterior credible interval based on the non-informative prior:

```r
a_noninformative/(a_noninformative + b_noninformative)
```

```
## [1] 0.3333333
```

```r
qbeta(c(0.025, 0.975), shape1 = a_noninformative, shape2 = b_noninformative)
```

```
## [1] 0.01257912 0.84188612
```

Prior mean and 95% posterior credible interval based on the weakly informative prior:

```r
a_weakly_informative/(a_weakly_informative + b_weakly_informative)
```

```
## [1] 0.1818182
```

```r
qbeta(c(0.025, 0.975), shape1 = a_weakly_informative, shape2 = b_weakly_informative)
```

```
## [1] 0.02521073 0.44501612
```

Prior mean and a 95% posterior credible interval based on the strongly informative prior:

```r
a_strongly_informative/(a_strongly_informative + b_strongly_informative)
```

```
## [1] 0.1980198
```

```r
qbeta(c(0.025, 0.975), shape1 = a_strongly_informative, shape2 = b_strongly_informative)
```

```
## [1] 0.1266556 0.2806980
```

Sample of size n = 10

The code below is exactly the same as the code above for exploring the prior distributions. Update the code to explore the posterior distributions obtained from each prior specification based on the data observed for your sample of size 10, and compare to the maximum likelihood estimate.

You will have to make two changes:

1. Update the a and b parameters to be the posterior parameters corresponding to each prior.
2. Add in a vertical line at the maximum likelihood estimate (use `geom_vline(xintercept = *)`, replacing the * with the maximum likelihood estimate).

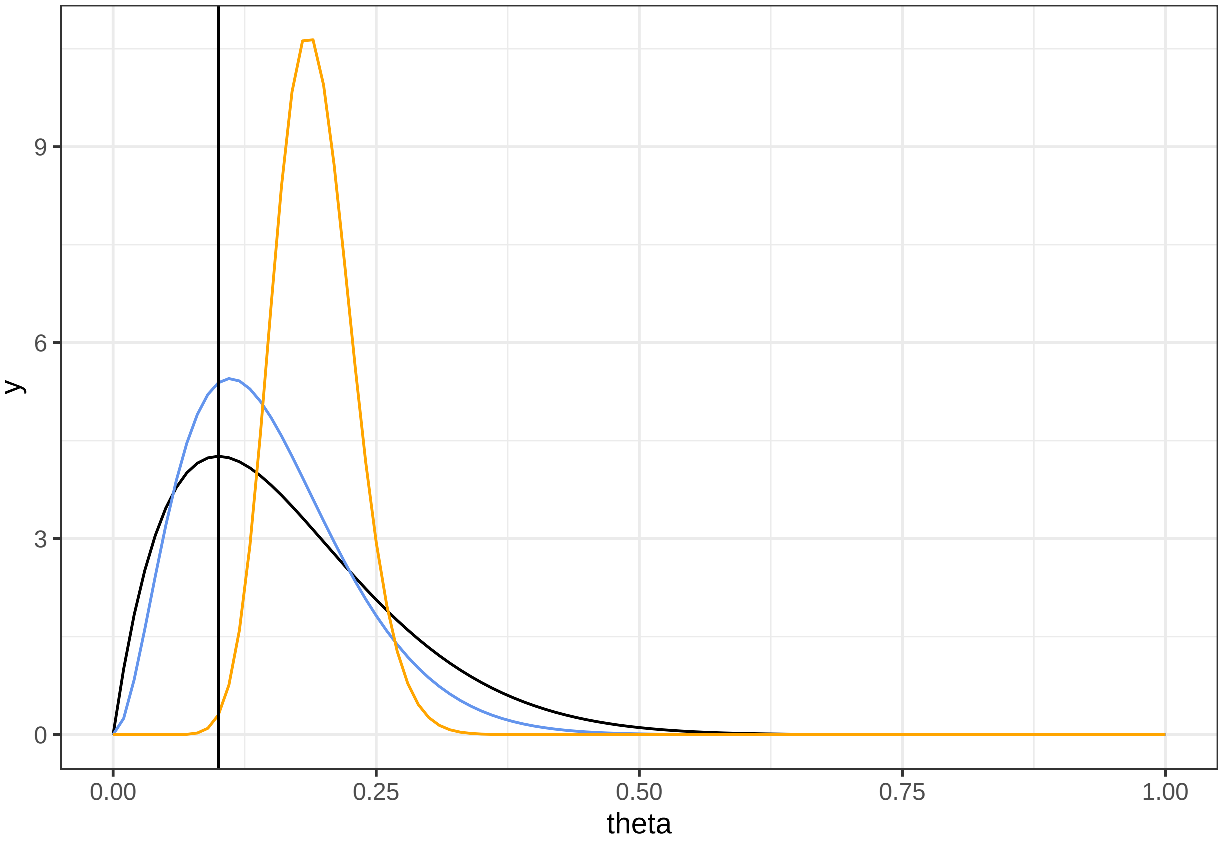
The rest of the code can be left as is.

```r
# I observed 1 blue in my first 10
a_noninformative <- 1 + 1
b_noninformative <- 1 + (10-1)

a_weakly_informative <- 2 + 1
b_weakly_informative <- 8 + (10-1)

a_strongly_informative <- 20 + 1
b_strongly_informative <- 80 + (10-1)

ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_noninformative, shape2 = b_noninformative)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_weakly_informative, shape2 = b_weakly_informative),
    color = "cornflowerblue") +
  stat_function(fun = dbeta,
    args = list(shape1 = a_strongly_informative, shape2 = b_strongly_informative),
    color = "orange") +
  geom_vline(xintercept = 1/10) +
  theme_bw()
```



Prior mean and 95% posterior credible interval based on the non-informative prior:

```
a_noninformative/(a_noninformative + b_noninformative)
```

```
## [1] 0.1666667
```

```
qbeta(c(0.025, 0.975), shape1 = a_noninformative, shape2 = b_noninformative)
```

```
## [1] 0.0228312 0.4127799
```

Prior mean and 95% posterior credible interval based on the weakly informative prior:

```
a_weakly_informative/(a_weakly_informative + b_weakly_informative)
```

```
## [1] 0.15
```

```
qbeta(c(0.025, 0.975), shape1 = a_weakly_informative, shape2 = b_weakly_informative)
```

```
## [1] 0.03382625 0.33137666
```

Prior mean and a 95% posterior credible interval based on the strongly informative prior:

```
a_strongly_informative/(a_strongly_informative + b_strongly_informative)
```

```
## [1] 0.1909091
```

```
qbeta(c(0.025, 0.975), shape1 = a_strongly_informative, shape2 = b_strongly_informative)
```

```
## [1] 0.1233955 0.2690554
```

Sample of size n = 20

The code below is exactly the same as the code above for exploring the prior distributions. Update the code to explore the posterior distributions obtained from each prior specification based on the data observed for your sample of size 20, and compare to the maximum likelihood estimate.

You will have to make two changes:

1. Update the a and b parameters to be the posterior parameters corresponding to each prior.
2. Add in a vertical line at the maximum likelihood estimate (use `geom_vline(xintercept = *)`, replacing the * with the maximum likelihood estimate).

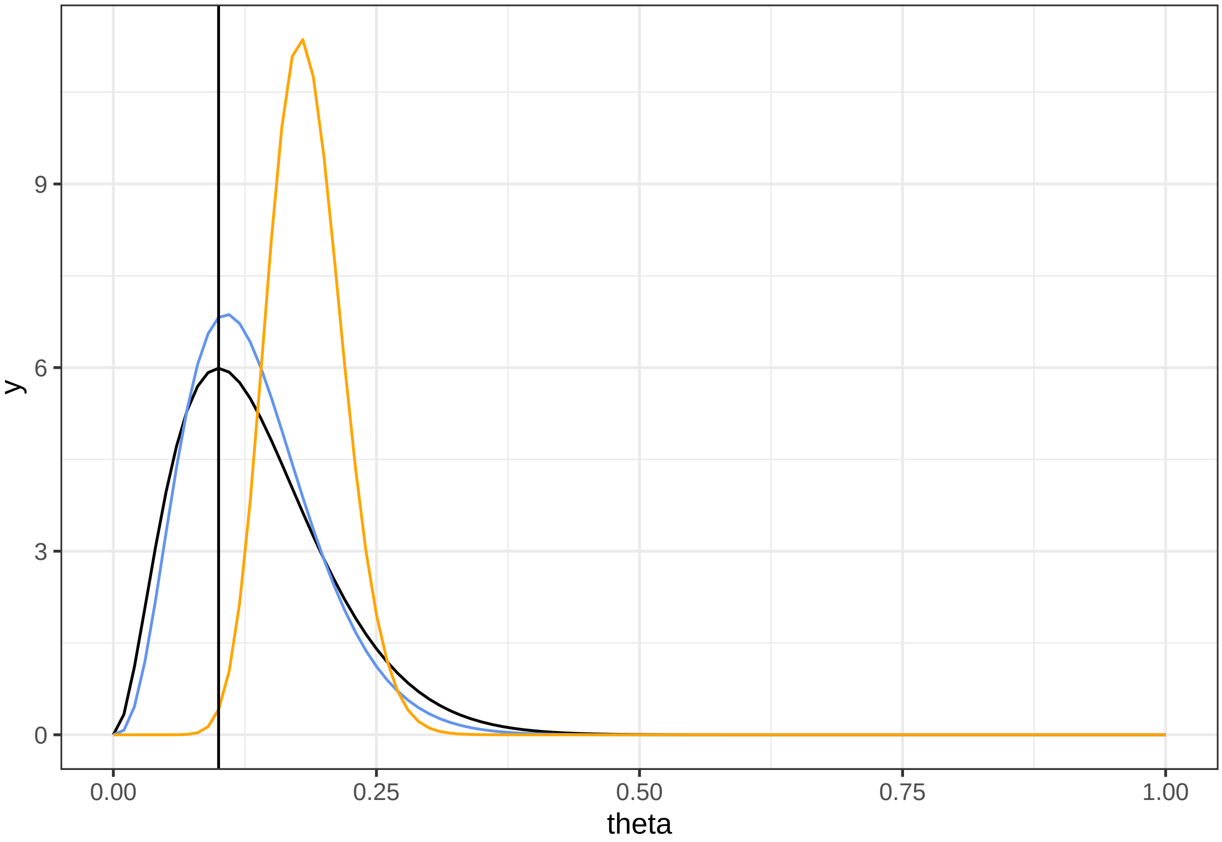
The rest of the code can be left as is.

```r
# I had 2 blue in my first 20
a_noninformative <- 1 + 2
b_noninformative <- 1 + (20 - 2)

a_weakly_informative <- 2 + 2
b_weakly_informative <- 8 + (20 - 2)

a_strongly_informative <- 20 + 2
b_strongly_informative <- 80 + (20 - 2)

ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_noninformative, shape2 = b_noninformative)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_weakly_informative, shape2 = b_weakly_informative),
    color = "cornflowerblue") +
  stat_function(fun = dbeta,
    args = list(shape1 = a_strongly_informative, shape2 = b_strongly_informative),
    color = "orange") +
  geom_vline(xintercept = 2/20) +
  theme_bw()
```



Prior mean and 95% posterior credible interval based on the non-informative prior:

```
a_noninformative/(a_noninformative + b_noninformative)
```

```
## [1] 0.1363636
```

```
qbeta(c(0.025, 0.975), shape1 = a_noninformative, shape2 = b_noninformative)
```

```
## [1] 0.03048897 0.30377441
```

Prior mean and 95% posterior credible interval based on the weakly informative prior:

```
a_weakly_informative/(a_weakly_informative + b_weakly_informative)
```

```
## [1] 0.1333333
```

```
qbeta(c(0.025, 0.975), shape1 = a_weakly_informative, shape2 = b_weakly_informative)
```

```
## [1] 0.03889483 0.27351520
```

Prior mean and a 95% posterior credible interval based on the strongly informative prior:

```
a_strongly_informative/(a_strongly_informative + b_strongly_informative)
```

```
## [1] 0.1833333
```

```
qbeta(c(0.025, 0.975), shape1 = a_strongly_informative, shape2 = b_strongly_informative)
```

```
## [1] 0.119637 0.257015
```

Sample of large size

The code below is exactly the same as the code above for exploring the prior distributions. Update the code to explore the posterior distributions obtained from each prior specification based on the data observed for your large sample size, and compare to the maximum likelihood estimate.

You will have to make two changes:

1. Update the a and b parameters to be the posterior parameters corresponding to each prior.
2. Add in a vertical line at the maximum likelihood estimate (use `geom_vline(xintercept = *)`, replacing the * with the maximum likelihood estimate).

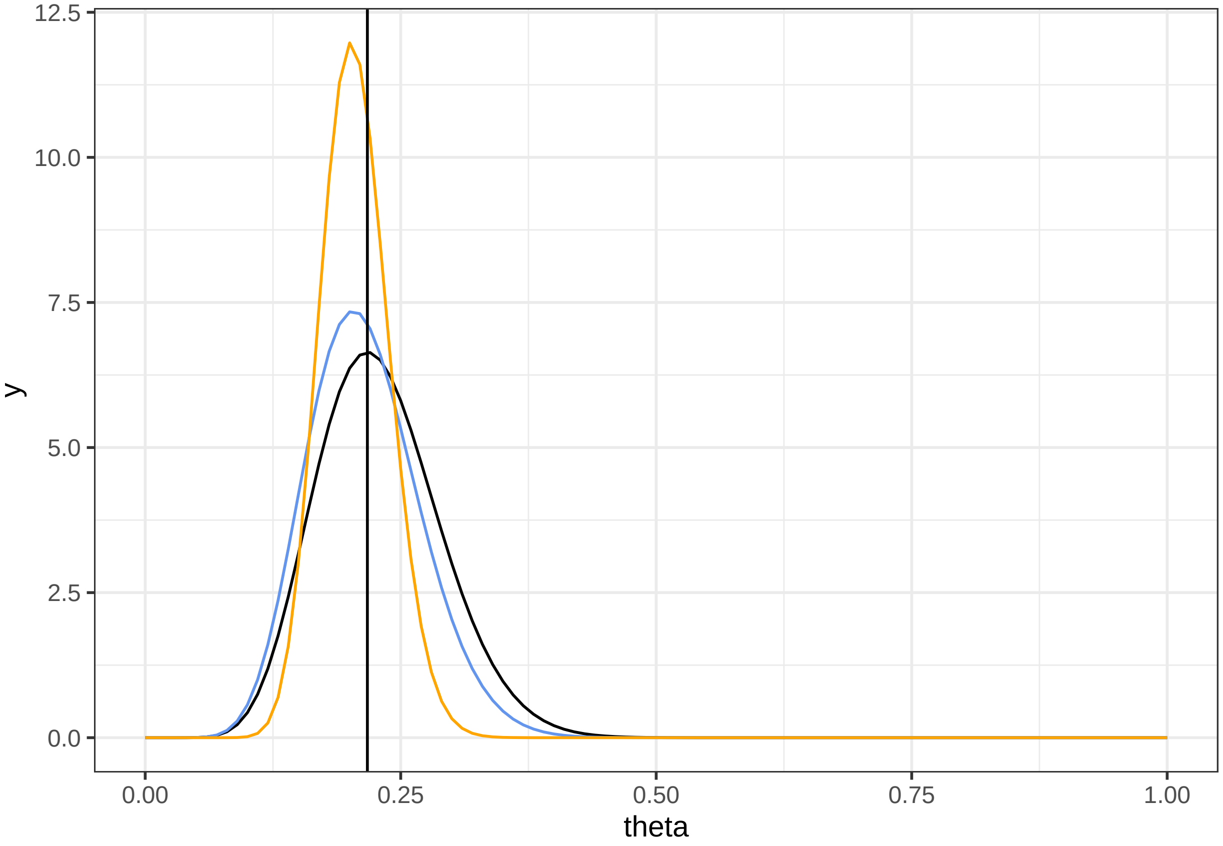
The rest of the code can be left as is.

```r
# I had 10 blue M&Ms in a sample of size 46
a_noninformative <- 1 + 10
b_noninformative <- 1 + (46 - 10)

a_weakly_informative <- 2 + 10
b_weakly_informative <- 8 + (46 - 10)

a_strongly_informative <- 20 + 10
b_strongly_informative <- 80 + (46 - 10)

ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_noninformative, shape2 = b_noninformative)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_weakly_informative, shape2 = b_weakly_informative),
    color = "cornflowerblue") +
  stat_function(fun = dbeta,
    args = list(shape1 = a_strongly_informative, shape2 = b_strongly_informative),
    color = "orange") +
  geom_vline(xintercept = 10/46) +
  theme_bw()
```



Prior mean and 95% posterior credible interval based on the non-informative prior:

```
a_noninformative/(a_noninformative + b_noninformative)
```

```
## [1] 0.2291667
```

```
qbeta(c(0.025, 0.975), shape1 = a_noninformative, shape2 = b_noninformative)
```

```
## [1] 0.1230335 0.3566370
```

Prior mean and 95% posterior credible interval based on the weakly informative prior:

```
a_weakly_informative/(a_weakly_informative + b_weakly_informative)
```

```
## [1] 0.2142857
```

```
qbeta(c(0.025, 0.975), shape1 = a_weakly_informative, shape2 = b_weakly_informative)
```

```
## [1] 0.1181358 0.3297285
```

Prior mean and a 95% posterior credible interval based on the strongly informative prior:

```
a_strongly_informative/(a_strongly_informative + b_strongly_informative)
```

```
## [1] 0.2054795
```

```
qbeta(c(0.025, 0.975), shape1 = a_strongly_informative, shape2 = b_strongly_informative)
```

```
## [1] 0.1441587 0.2744374
```

Combined sample across everyone in class

The code below is exactly the same as the code above for exploring the prior distributions. Update the code to explore the posterior distributions obtained from each prior specification based on the data observed for your large sample size, and compare to the maximum likelihood estimate.

You will have to make two changes:

1. Update the a and b parameters to be the posterior parameters corresponding to each prior.
2. Add in a vertical line at the maximum likelihood estimate (use `geom_vline(xintercept = *)`, replacing the * with the maximum likelihood estimate).

The rest of the code can be left as is.

```
# We had the following overall count of blue M&Ms and sample size:
x <- 4 + 10 + 7 + 9 + 12 + 12 + 18 + 7 + 16 + 16
n <- 58 + 47 + 53 + 59 + 49 + 53 + 75 + 43 + 72 + 50
x
```

```
## [1] 111
```

```
n
```

```
## [1] 559
```

```
x/n
```

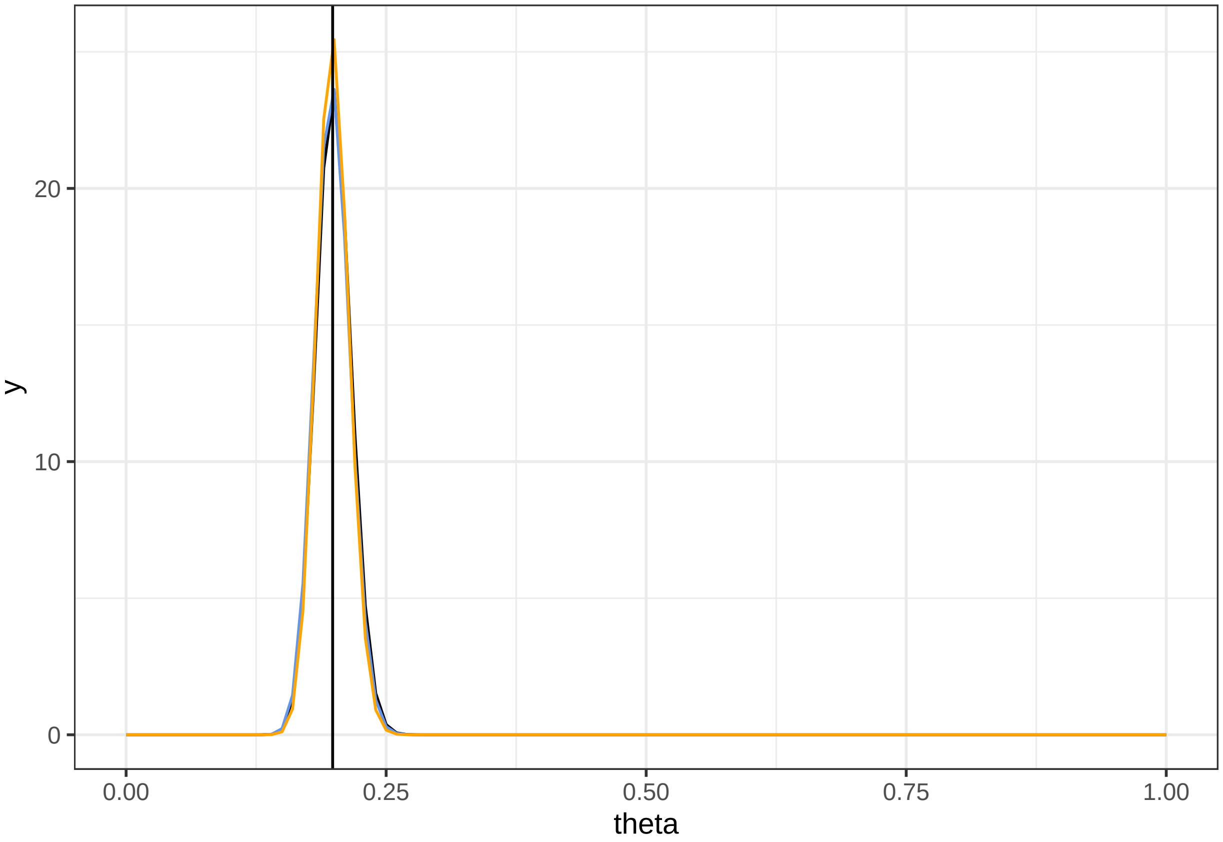
```
## [1] 0.1985689
```

```
a_noninformative <- 1 + x
b_noninformative <- 1 + (n - x)

a_weakly_informative <- 2 + x
b_weakly_informative <- 8 + (n - x)

a_strongly_informative <- 20 + x
b_strongly_informative <- 80 + (n - x)

ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_noninformative, shape2 = b_noninformative)) +
  stat_function(fun = dbeta,
    args = list(shape1 = a_weakly_informative, shape2 = b_weakly_informative),
    color = "cornflowerblue") +
  stat_function(fun = dbeta,
    args = list(shape1 = a_strongly_informative, shape2 = b_strongly_informative),
    color = "orange") +
  geom_vline(xintercept = x/n) +
  theme_bw()
```



Prior mean and 95% posterior credible interval based on the non-informative prior:

```
a_noninformative/(a_noninformative + b_noninformative)
```

```
## [1] 0.1996435
```

```
qbeta(c(0.025, 0.975), shape1 = a_noninformative, shape2 = b_noninformative)
```

```
## [1] 0.1676333 0.2336817
```

Prior mean and 95% posterior credible interval based on the weakly informative prior:

```
a_weakly_informative/(a_weakly_informative + b_weakly_informative)
```

```
## [1] 0.198594
```

```
qbeta(c(0.025, 0.975), shape1 = a_weakly_informative, shape2 = b_weakly_informative)
```

```
## [1] 0.1668702 0.2323243
```

Prior mean and a 95% posterior credible interval based on the strongly informative prior:

```
a_strongly_informative/(a_strongly_informative + b_strongly_informative)
```

```
## [1] 0.198786
```

```
qbeta(c(0.025, 0.975), shape1 = a_strongly_informative, shape2 = b_strongly_informative)
```

```
## [1] 0.1692238 0.2300797
```

Understanding what you saw above

Pick one sample size (maybe 10). How do the three posterior distributions for θ compare? How does the strength of prior knowledge relate to the strength of posterior knowledge? How do the three point estimates and the credible intervals compare? Are some intervals wider or narrower than others?

With small sample sizes, the three posterior distributions are fairly different and point estimates (posterior means) are fairly different from the maximum likelihood estimate. When the sample size is small, the posterior distribution based on a strongly informative prior doesn't change much relative to the prior. For the weaker priors, the posterior distribution changes more based on the data when the sample size is small. The posterior credible intervals based on the strongly informative prior are narrower than the posterior credible intervals based on the weakly informative priors.

Focus now on how the posterior distribution coming out of the analysis using a weakly informative prior distribution changes as the sample size increases. How do the posterior means compare to the maximum likelihood estimate? How does the width of the posterior credible interval change as the sample size increases?

When the sample size is large, all posterior distributions agree pretty closely with each other, all three posterior credible intervals are very similar, and all three posterior means are similar to the maximum likelihood estimate. The posterior credible intervals are narrower when the sample size is larger.