S626

HW2

John Koo

```r
library(ggplot2)

theme_set(theme_bw())
```

3.3

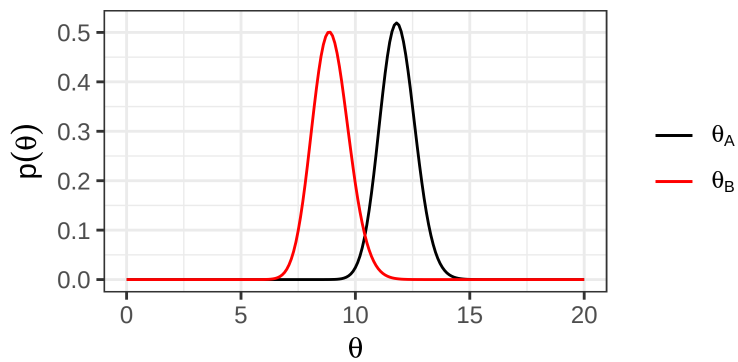
```r
# data
y.a <- c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
y.b <- c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)

# statistics
n.a <- length(y.a)
n.b <- length(y.b)
sum.y.a <- sum(y.a)
sum.y.b <- sum(y.b)
```

a

From class, we know:

- $\theta_A \mid y_A \sim Gamma(120 + \sum y_{A,i}, 10 + n_A)$
  $\implies \theta_A \mid y_A \sim Gamma(237, 20)$
- $\theta_B \mid y_B \sim Gamma(12 + \sum y_{B,i}, 1 + n_B) \implies \theta_B \mid y_B \sim Gamma(125, 14)$

```r
# prior parameters
a.a <- 120
a.b <- 10
b.a <- 12
b.b <- 1

# posterior distributions
theta.space <- seq(0, 20, .1)
theta.a <- dgamma(theta.space, a.a + sum.y.a, a.b + n.a)
theta.b <- dgamma(theta.space, b.a + sum.y.b, b.b + n.b)

ggplot() +
  geom_line(aes(x = theta.space, y = theta.a, colour = 'theta[A]')) +
  geom_line(aes(x = theta.space, y = theta.b, colour = 'theta[B]')) +
  scale_colour_manual(labels = c(expression(theta[A]), expression(theta[B])),
                      values = c(1, 2),
                      name = NULL) +
  labs(x = expression(theta), y = expression(p(theta)))
```



$E[\theta_A|y_A] = Var(\theta_A|y_A) = \frac{120+\sum y_{A,i}}{10+n_A} = 11.85$

$E[\theta_B|y_B] = Var(\theta_B|y_B) = \frac{12+\sum y_{B,i}}{1+n_B} = 8.9286$

For the 95% credible intervals:

```
alpha <- .05

# strain A
c('lower' = qgamma(alpha / 2, a.a + sum.y.a, a.b + n.a),
  'upper' = qgamma(1 - alpha / 2, a.a + sum.y.a, a.b + n.a))
```

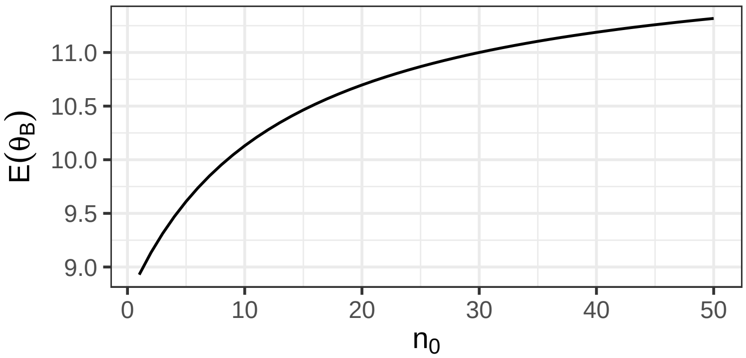
```
   lower    upper
10.38924 13.40545
```

```
# strain B
c('lower' = qgamma(alpha / 2, b.a + sum.y.b, b.b + n.b),
  'upper' = qgamma(1 - alpha / 2, b.a + sum.y.b, b.b + n.b))
```

```
    lower     upper
 7.432064 10.560308
```

b

```
n.0 <- seq(50)
mean.theta.b <- (b.a * n.0 + sum.y.b) / (n.0 + n.b)

ggplot() +
  geom_line(aes(x = n.0, y = mean.theta.b)) +
  labs(x = expression(n[0]),
       y = expression(E(theta[B])))
```



The MLE for θ is just the sample mean. The sample mean for strain B is close to 9, while the posterior mean for strain A is close to 12. So we would need a large n_0 to make the prior dominate the sample.

c

We would expect some relationship between the two groups. Perhaps something we can say is in addition to $\theta_A \sim Gamma(a_A, b_A)$ and $\theta_B \sim Gamma(a_B, b_B)$, we can say that the parameters of these gamma distributions also come from some (shared) distribution.

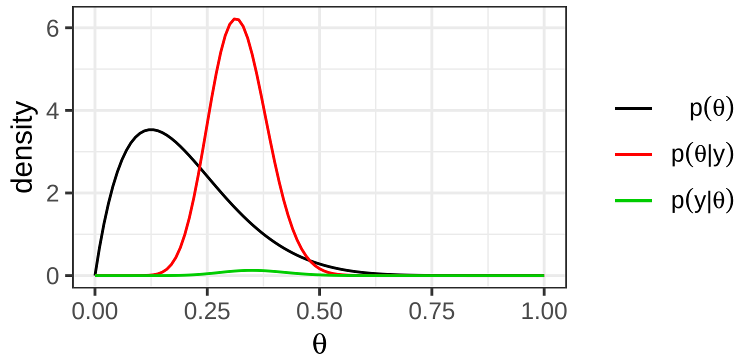
3.4

a

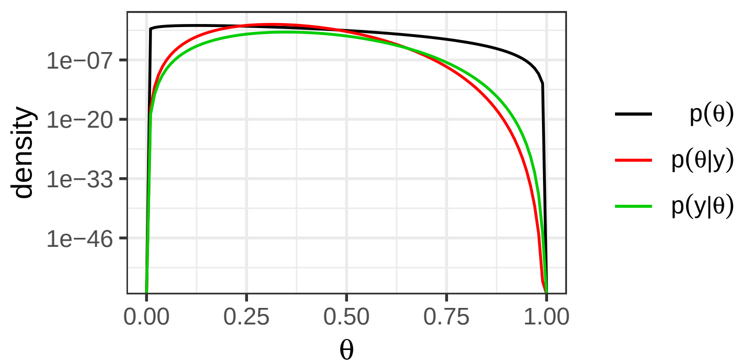
From class, we know that $\theta \mid y \sim Beta(y + a, n - y + b)$.

```
# data
n <- 43
y <- 15

# parameters
a <- 2
b <- 8

# support
theta.vector <- seq(0, 1, .01)

# densities
p.theta <- dbeta(theta.vector, a, b)
p.theta.y <- dbeta(theta.vector, y + a, n - y + b)
p.y.theta <- dbinom(y, n, theta.vector)

# plot
p.3.4.a.plot <- ggplot() +
  geom_line(aes(x = theta.vector, y = p.theta, colour = '1')) +
  geom_line(aes(x = theta.vector, y = p.theta.y, colour = '2')) +
  geom_line(aes(x = theta.vector, y = p.y.theta, colour = '3')) +
  scale_colour_manual(labels = c(expression(p(theta)),
                                 expression(p(theta*'|'*y)),
                                 expression(p(y*'|'*theta))),
```

```r
                    values = seq(3),
                    name = NULL) +
  labs(y = 'density', x = expression(theta))

p.3.4.a.plot
```




```r
p.3.4.a.plot + scale_y_log10()
```



For the mean and variance (and standard deviation), for some $X \sim Beta(a, b)$,

- $E[X] = \int_0^1 x \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}dx$
  $= \frac{1}{B(a,b)} \int x^{a+1-1}(1-x)^{b-1}dx$
  $= \frac{B(a+1,b)}{B(a,b)} = \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$
  $= \frac{a}{a+b}$
- Similarly, $E[X^2] = \frac{(a+1)a}{(a+b+1)(a+b)}$,
  so $Var(X) = E[X^2] - (E[X])^2 = \frac{(a+1)a}{(a+b+1)(a+b)} - \frac{a^2}{(a+b)^2}$
  $= \frac{a(a^2+a+ab+b-a^2-ab-a)}{(a+b)^2(a+b+1)}$
  $= \frac{ab}{(a+b)^2(a+b+1)}$
- For the mode, we need $\sup_x \frac{1}{B(a,b} x^{a-1}(1-x)^{b-1}$
  which we can solve by differentiating once, setting to 0, and solving for x, and we obtain the following equation:
  $0 = \frac{x^{a-2}(1-x)^{b-1}}{a-1} - \frac{x^{a-1}(1-x)^{b-2}}{b-1}$
  $\implies 0 = (b-1)x - (1-x)(a-1) = (a+b-2)x - a + 1$
  $\implies x = \frac{a-1}{a+b-2}$

$E[\theta|y] = \frac{y+a}{y+a+n-y+b} \approx 0.3208$

$mode(\theta|y) = \frac{y+a-1}{y+a+n-y+b-2} \approx 0.3137$

$sd(\theta|y) = \sqrt{\frac{(y+a)(n-y+b)}{(y+a+n-y+b)^2(y+a+n-y+b-1)}} \approx 0.0647$

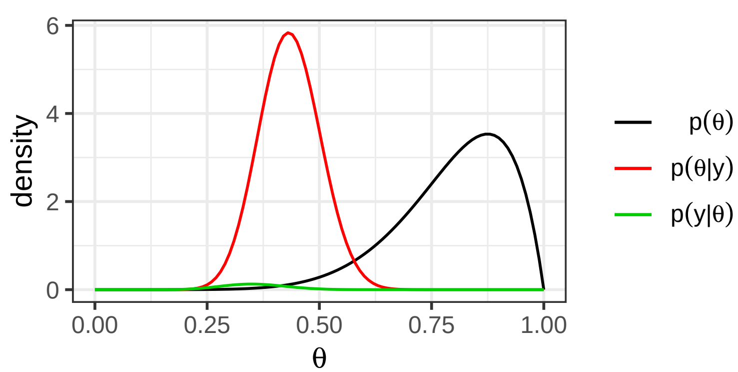
95% CI:

```
alpha <- .05
c('lower' = qbeta(alpha / 2, y + a, n - y + b),
  'upper' = qbeta(1 - alpha / 2, y + a, n - y + b))
```

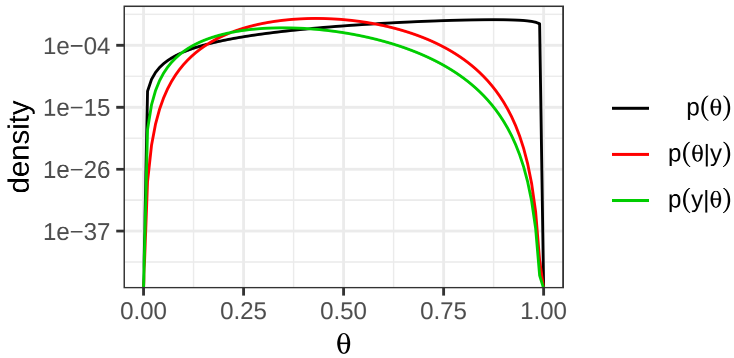
```
    lower     upper
0.2032978 0.4510240
```

b

```
# parameters
a <- 8
b <- 2

# densities
p.theta <- dbeta(theta.vector, a, b)
p.theta.y <- dbeta(theta.vector, y + a, n - y + b)
p.y.theta <- dbinom(y, n, theta.vector)

# plot
p.3.4.b.plot <- ggplot() +
  geom_line(aes(x = theta.vector, y = p.theta, colour = '1')) +
  geom_line(aes(x = theta.vector, y = p.theta.y, colour = '2')) +
  geom_line(aes(x = theta.vector, y = p.y.theta, colour = '3')) +
  scale_colour_manual(labels = c(expression(p(theta)),
                                 expression(p(theta*'|'*y)),
                                 expression(p(y*'|'*theta))),
                      values = seq(3),
                      name = NULL) +
  labs(y = 'density', x = expression(theta))

p.3.4.b.plot
```




```
p.3.4.b.plot + scale_y_log10()
```



$E[\theta|y] = \frac{y+a}{y+a+n-y+b} \approx 0.434$

$mode(\theta|y) = \frac{y+a-1}{y+a+n-y+b-2} \approx 0.4314$

$sd(\theta|y) = \sqrt{\frac{(y+a)(n-y+b)}{(y+a+n-y+b)^2(y+a+n-y+b-1)}} \approx 0.0687$

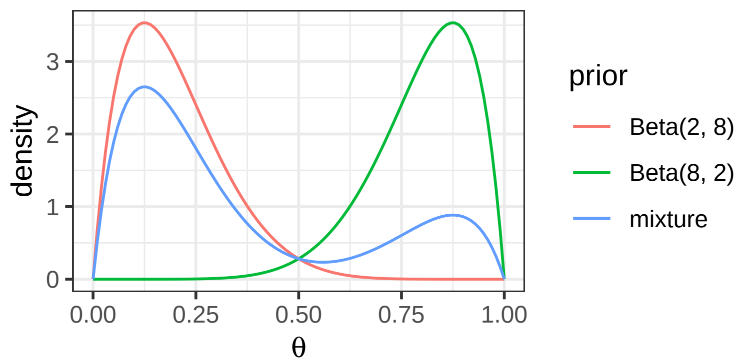
95% CI:

```
alpha <- .05
c('lower' = qbeta(alpha / 2, y + a, n - y + b),
  'upper' = qbeta(1 - alpha / 2, y + a, n - y + b))
```

```
    lower     upper
0.3046956 0.5679528
```

c

```
p1 <- dbeta(theta.vector, 2, 8)
p2 <- dbeta(theta.vector, 8, 2)
p3 <-
  .25 * gamma(10) / gamma(2) / gamma(8) *
  (3 * theta.vector * (1 - theta.vector) ** 7 +
     theta.vector ** 7 * (1 - theta.vector))

ggplot() +
  geom_line(aes(x = theta.vector, y = p1, colour = '1')) +
  geom_line(aes(x = theta.vector, y = p2, colour = '2')) +
  geom_line(aes(x = theta.vector, y = p3, colour = '3')) +
  scale_colour_discrete(labels = c('Beta(2, 8)', 'Beta(8, 2)', 'mixture'),
                        name = 'prior') +
  labs(x = expression(theta), y = 'density')
```

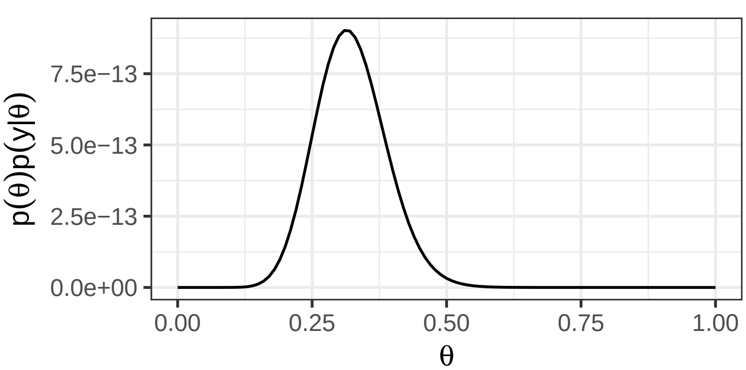


This might be a good choice if we have some evidence that θ should be either $1/5$ or $4/5$.

d

$\frac{1}{4}\frac{\Gamma(10)}{\Gamma(2)\Gamma(8)}(3\theta(1-\theta)^7\theta^7(1-\theta))(\binom{n}{y}\theta^y(1-\theta)^{n-y})$
$\propto (3\theta(1-\theta)^7\theta^7(1-\theta))\theta^y(1-\theta)^{n-y}$
$= 3\theta^{y+2-1}(1-\theta)^{n-y+8-1} + \theta^{y+8-1}(1-\theta)^{n-y+2-1}$
$\implies$ θ is a mixture of $Beta(y+2, n-y+8)$ and $Beta(y+8, n-y+2)$ distributions.

```
.p <- function(theta) {
  .25 * gamma(10) / gamma(2) / gamma(8) *
    (3 * theta ** (y + 1) * (1 - theta) ** (n - y + 7) +
       theta ** (y + 7) * (1 - theta) ** (n - y + 1))
}

density.vector <- sapply(theta.vector, .p)

ggplot() +
  geom_line(aes(x = theta.vector, y = density.vector)) +
  labs(x = expression(theta),
       y = expression(p(theta) * p(y*'|'*theta)))
```




```
# posterior mode approximation
optimize(.p, interval = c(0, 1), maximum = TRUE)
```

```
$maximum
[1] 0.3140734

$objective
```

```
[1] 9.035285e-13
```

e

We can say $\theta \mid a_1, a_2, n, q \sim p(\theta|a_1, a_2, n, q)$ where $p(\theta|a, b, q) = \frac{\Gamma(n)}{\Gamma(a_1)\Gamma(a_2)}(q \times Beta(a_1, n - a_1) + (1-q) \times Beta(a_2, n - a_2))$
where $a_1, a_2 < n$.

- n is the prior sample size
- We have some reason to believe that θ should either be a_1/n or a_2/n
- q is how confident we are that $\theta = a_1/n$ instead of a_2/n
- $1 - q$ is how confident we are that $\theta = a_2/n$ instead of a_1/n

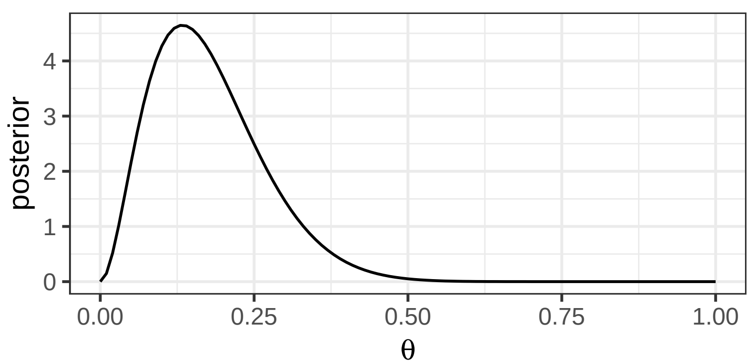
3.7

a

From class, we know that $\theta \sim Beta(y_1 + 1, n_1 - y_1 + 1) = Beta(3, 14)$. Similar to the previous problem, we know:

- $E[\theta|y_1] = \frac{3}{3+14} \approx 0.1765$
- $sd(\theta|y_1) = \sqrt{\frac{(3)(14)}{(3+14)^2(14-1)}} \approx 0.1057$
- $mode(\theta|y) = \frac{2}{15} \approx 0.1333$

```
y.1 <- 2
n.1 <- 15

p <- dbeta(theta.vector, y.1 + 1, n.1 - y.1 + 1)

ggplot() +
  geom_line(aes(x = theta.vector, y = p)) +
  labs(x = expression(theta), y = 'posterior')
```



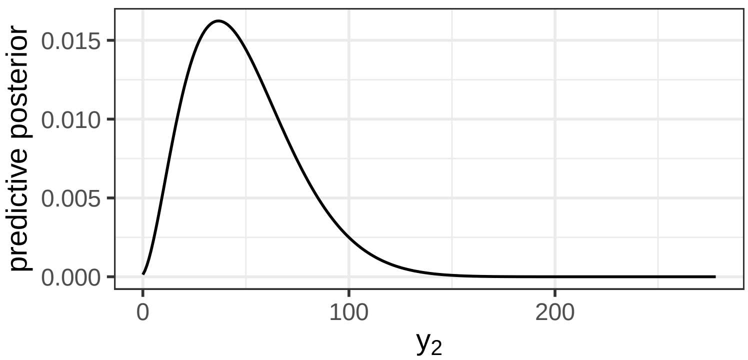
b

$p(y_2|y_1) = \int p(y_2|y_1, \theta)p(\theta|y_1)d\theta$, so here we are saying $p(y_2|y_1, \theta) = p(y_2, \theta)$, i.e., $\tilde{Y} \perp Y \mid \theta$, i.e., each observation is independent of the others.

$p(y_2|y_1) = \int_0^1 p(y_2|\theta)p(\theta|y_1)d\theta$
$= \int \binom{n_2}{y_2}\theta^{y_2}(1-\theta)^{n_2-y_2}\frac{1}{B(y_1+1,n_1-y_1+1)}\theta^{y_1}(1-\theta)^{n_1-y_1}d\theta$
$= \frac{\binom{n_2}{y_2}}{B(y_1+1,n_1-y_1+1)}\int \theta^{y_2+y_1}(1-\theta)^{n_2+n_1-y_2-y_1}d\theta$
$= \binom{n_2}{y_2}\frac{B(y_1+1,n_1-y_1+1)}{B(y_2+y_1+1,n_2+n_1-y_2-y_1+1)}$

c

```
n.2 <- 278
y.2 <- seq(0, n.2)

p <- choose(n.2, y.2) / beta(y.1 + 1, n.1 - y.1 + 1) *
  beta(y.1 + y.2 + 1, n.2 + n.1 - y.2 - y.1 + 1)

ggplot() +
  geom_line(aes(x = y.2, y = p)) +
  labs(x = expression(y[2]), y = 'predictive posterior')
```



$E[Y_2|Y_1] = E[Y_2|(\theta|Y_1)] = n_2 \times \frac{y_1+1}{n_1+2} \approx 49.0588$

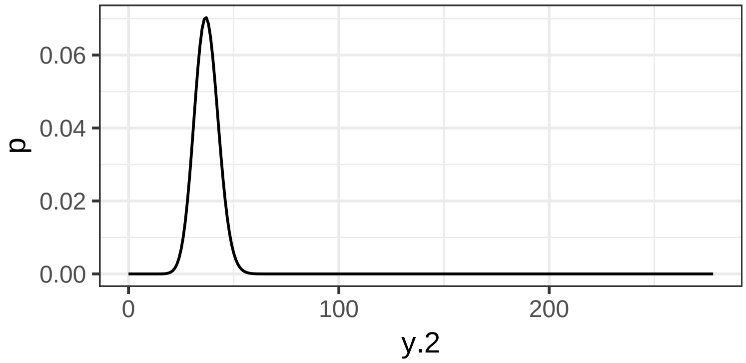
$Var(Y_2|Y_1) = E[Var(Y_2|(\theta|Y_1))] + Var(E[Y_2|(\theta|Y_1)])$
$= E[n_2\theta(1-\theta)|Y_1] + Var(n_2\theta|Y_1)$
$= n_1(\frac{y_1+1}{n_1+2} - \frac{(y_1+2)(y_1+1)}{(n_1+3)(n_1+2)}) + n_2^2\frac{(y_1+1)(n_1-y_1+1)}{(n_1+2)^2(n_1+3)}$
And $sd(Y_2|Y_1) = \sqrt{Var(Y_2|Y_1)}$

```
posterior.var <- n.1 *
  ((y.1 + 1) / (n.1 + 2) - (y.1 + 2) *
     (y.1 + 1) / (n.1 + 3) / (n.1 + 2)) / (n.1 + 2) +
  n.2 ** 2 * (y.1 + 1) * (n.1 - y.1 + 1) / (n.1 + 2) ** 2 / (n.1 + 3)
sqrt(posterior.var)
```

```
[1] 24.98195
```

d

```
p.hat <- y.1 / n.1
p <- dbinom(y.2, n.2, p.hat)

ggplot() +
  geom_line(aes(x = y.2, y = p))
```




```
# expected value
p.hat * n.2
```

```
[1] 37.06667
```

```
# standard deviation
sqrt(p.hat * (1 - p.hat) * n.2)
```

```
[1] 5.667843
```

Since the first experiment has a small sample size, the prior has a large effect. Compare $y_1/n_1 \approx 0.1333$ vs $\frac{y_1+1}{n_1+2} \approx 0.1765$.

The standard deviation is much smaller since we are artificially inflating our knowledge of θ.