STA 602 Lab

Student

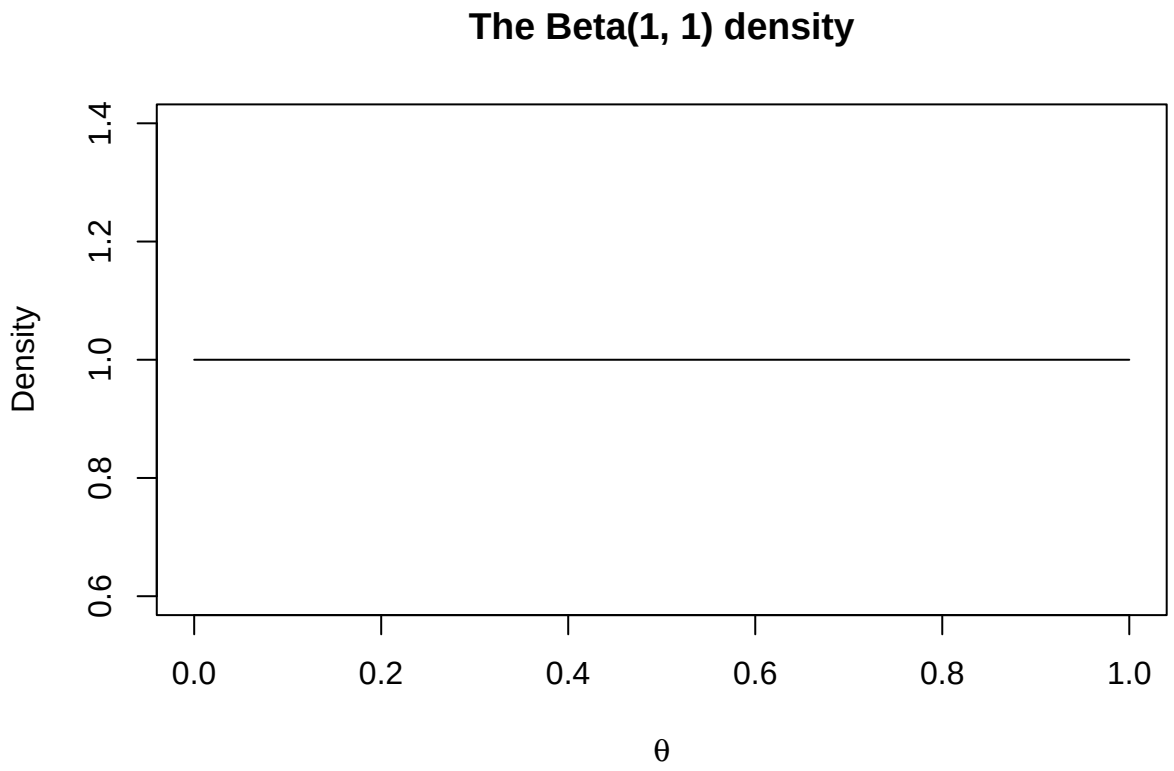
12 September, 2022

Exercise 1

The histogram below is unimodal and approximately normal with mean around .155.

```
tumors <- read.csv(file = url("http://www.stat.columbia.edu/~gelman/book/data/rats.asc"),
                   skip = 2, header = T, sep = " ")[,c(1,2)]
y <- tumors$y
N <- tumors$N
n <- length(y)


plot(seq(0, 1, length.out = 1000),
     dbeta(seq(0, 1, length.out = 1000), 1, 1),
     type = 'l',
     xlab = expression(theta), ylab = "Density",
     main = "The Beta(1, 1) density")
```




```
stan_dat <- list(n = n, N = N, y =y, a = 1, b = 1)
fit_pool <- stan('lab-02-pool.stan', data = stan_dat, chains = 2, refresh = 0)
```

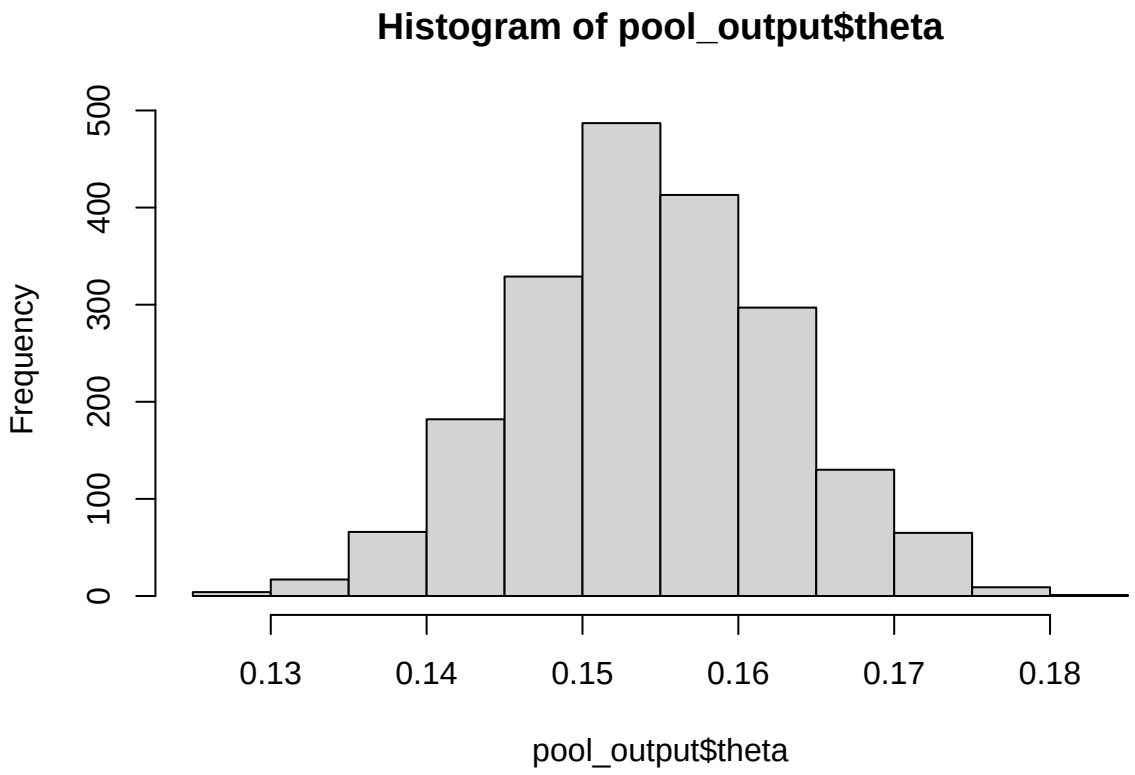
```
## Trying to compile a simple C file
```

```
## Running /usr/lib64/R/bin/R CMD SHLIB foo.c
## gcc -m64 -I"/usr/include/R" -DNDEBUG   -I"/usr/lib64/R/library/Rcpp/include/"  -I"/usr/lib64/R/libra
## In file included from /usr/lib64/R/library/RcppEigen/include/Eigen/Dense:1,
##                  from /usr/lib64/R/library/StanHeaders/include/stan/math/prim/mat/fun/Eigen.hpp:13,
##                  from <command-line>:
## /usr/lib64/R/library/RcppEigen/include/Eigen/Core:82:12: fatal error: new: No such file or directory
##    82 |   #include <new>
##       |            ^~~~~
## compilation terminated.
## make: *** [/usr/lib64/R/etc/Makeconf:168: foo.o] Error 1
```

```
pool_output <- rstan::extract(fit_pool)
mean(pool_output$theta)
```

```
## [1] 0.1543446
```

```
hist(pool_output$theta)
```



Exercise 2

each point on the x axis is a column, representing the different labs, and each point plotted represents different samples of the theta_i

```
stan_dat <- list(n = n, N = N, y =y, a = 1, b = 1)
fit_nopool <- stan('lab-02-nopool.stan', data = stan_dat, chains = 2, refresh = 0)
```

```
## Trying to compile a simple C file
```

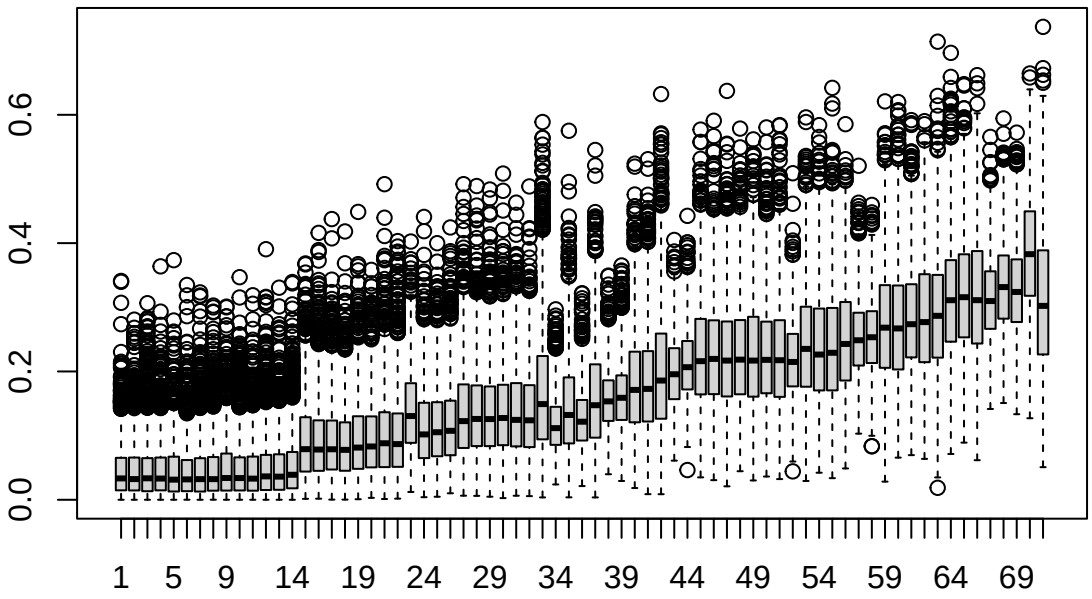
```
## Running /usr/lib64/R/bin/R CMD SHLIB foo.c
## gcc -m64 -I"/usr/include/R" -DNDEBUG   -I"/usr/lib64/R/library/Rcpp/include/"  -I"/usr/lib64/R/libra
```

```
## In file included from /usr/lib64/R/library/RcppEigen/include/Eigen/Dense:1,
##                  from /usr/lib64/R/library/StanHeaders/include/stan/math/prim/mat/fun/Eigen.hpp:13,
##                  from <command-line>:
## /usr/lib64/R/library/RcppEigen/include/Eigen/Core:82:12: fatal error: new: No such file or directory
##    82 |   #include <new>
##       |            ^~~~~
## compilation terminated.
## make: *** [/usr/lib64/R/etc/Makeconf:168: foo.o] Error 1
```

```
nopool_output <- rstan::extract(fit_nopool)
apply(nopool_output$theta,2,mean)
```

```
##  [1] 0.04567157 0.04576556 0.04660192 0.04626698 0.04571170 0.04391771
##  [7] 0.04575983 0.04654547 0.04886813 0.04818146 0.04693485 0.05020337
## [13] 0.04926412 0.05341878 0.09208530 0.09115428 0.09156452 0.08919683
## [19] 0.09532936 0.09568669 0.10048229 0.09996463 0.13931234 0.11332447
## [25] 0.11519755 0.11896966 0.13565440 0.13705860 0.13649618 0.13787223
## [31] 0.13720955 0.13556533 0.16700761 0.11711174 0.14425624 0.12640464
## [37] 0.15896685 0.15713905 0.16201851 0.18103734 0.18220362 0.19938092
## [43] 0.19914791 0.21155529 0.22683377 0.22704667 0.22413363 0.22667050
## [49] 0.22611509 0.22643210 0.22533897 0.21967675 0.24164775 0.23747238
## [55] 0.23871858 0.24989924 0.25190514 0.25527928 0.27311459 0.27401689
## [61] 0.28174911 0.28648723 0.29209355 0.31616697 0.31959054 0.31830428
## [67] 0.31217633 0.33338331 0.32673493 0.38528495 0.31192965
```

```
boxplot(nopool_output$theta)
```



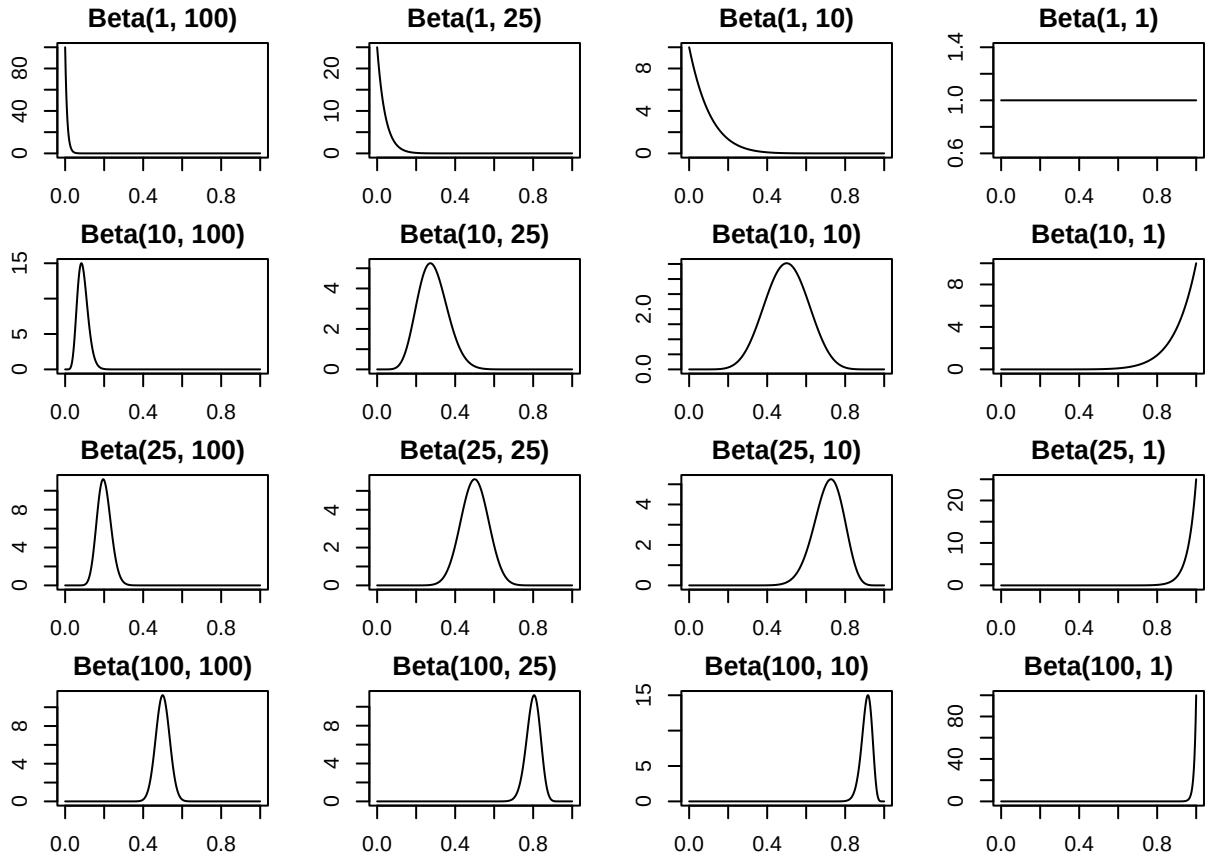
Exercise 3

I think the difference is that the pool file uses one random variable for all thetas, given by the "real" keyword i.e. all of the different labs are in the same pool. The nopool file uses different thetas, represented by a vector of thetas, for each lab.

Exercise 4

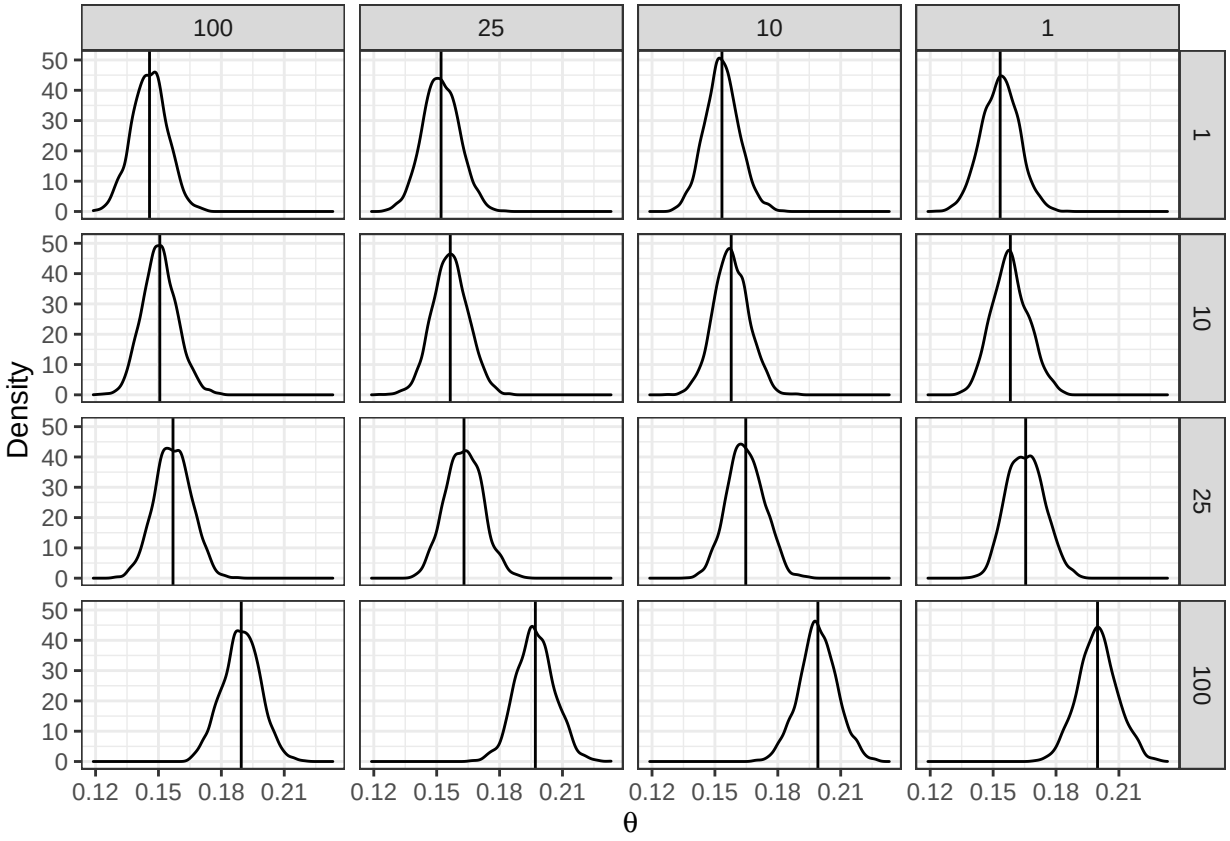
Each a represents our prior belief about the number of tumors we believe will be found in a sample of rats and b represents the number of "failures", or rats without tumors, we observe. This could be informed based on past studies, or if we have no idea, could be represented with the beta(1,1) prior.

```
#plotting different priors
par(mfrow = c(4, 4))
par(mar=c(2,2,2,2))
for(a_val in c(1, 10, 25, 100)){
  for(b_val in rev(c(1, 10, 25, 100))){
    plot(seq(0, 1, length.out = 1000),
     dbeta(seq(0, 1, length.out = 1000), a_val, b_val),
     type = 'l',
     xlab = expression(theta), ylab = "Density",
     main = paste0("Beta(", a_val, ", ", b_val, ")"))
  }
}
```




```
#getting samples from each posterior given the above priors
output_list <- list()
for(a_val in c(1, 10, 25, 100)){
  for(b_val in c(1, 10, 25, 100)){
    stan_dat <- list(n = n, N = N, y = y, a = a_val, b = b_val)
    fit_pool <- stan('lab-02-pool.stan', data = stan_dat, chains = 2, refresh = 0)
    output_list[[paste0("a_", a_val, ":b_", b_val)]] <- rstan::extract(fit_pool)[["theta"]]
  }
}
```

```r
#compiling above samples into a dataframe
output_list %>%
  plyr::ldply(function(theta){
    reshape2::melt(theta) %>%
      dplyr::mutate(post_mean = mean(theta))
  }, .id = "prior") %>%
  tidyr::separate("prior", into = c("a", "b"), sep = ":") %>%
  dplyr::mutate(a = as.numeric(gsub("._", "", a)),
                b = as.numeric(gsub("._", "", b))) %>%
  ggplot2::ggplot() +
  geom_density(aes(x = value)) +
  geom_vline(aes(xintercept = post_mean)) +
  facet_grid(a~factor(b, levels = rev(c(1, 10, 25, 100)))) +
  scale_colour_brewer(palette = "Set1") +
  labs(x = expression(theta), y = "Density")
```



Exercise 5

From the beginning of the lab, "specifically, we have the number of incidences of endometrial stromal polyps in 71 different groups of female lab rats of type F344". So each y is the actual observed number of successes, a, and, N-a, would be the value for b, or the "failures".

Exercise 6

Below is a calculation of the MLE's, Y-bar of the different observed tumors from the N labs. It looks like our priors beliefs range from about .15 to .20 which is a pretty narrow range, while the MLEs below range from 0

to 0.375 which is a much wider range clearly.

```
mle = tumors$y / tumors$N
max(mle)
```

```
## [1] 0.375
```

```
min(mle)
```

```
## [1] 0
```

Exercise 7

Why might we have observed such a difference between the two approaches when using the prior Beta(1,1)? Consider calculating the MLEs for theta and theta_i and comparing these values to the values obtained with the Bayesian approach:

If we feed the values below into our modified formula for E[theta|Y}, allowing the MLE to vary widely will result in different expectations of theta. This will then change the values of the Bayesian approach.

```
# approach 1
mle.1 <- sum(y)/sum(N)

# approach 2
mle.2 <- y/N

mle.1
```

```
## [1] 0.1535365
```

```
mle.2
```

```
##  [1] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##  [7] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## [13] 0.00000000 0.00000000 0.05000000 0.05000000 0.05000000 0.05000000
## [19] 0.05263158 0.05263158 0.05555556 0.05555556 0.11111111 0.08000000
## [25] 0.08333333 0.08695652 0.10000000 0.10000000 0.10000000 0.10000000
## [31] 0.10000000 0.10000000 0.10000000 0.10204082 0.10526316 0.10869565
## [37] 0.11764706 0.14285714 0.14893617 0.15000000 0.15000000 0.15384615
## [43] 0.18750000 0.20000000 0.20000000 0.20000000 0.20000000 0.20000000
## [49] 0.20000000 0.20000000 0.20000000 0.20833333 0.21052632 0.21052632
## [55] 0.21052632 0.22727273 0.23913043 0.24489796 0.25000000 0.25000000
## [61] 0.26086957 0.26315789 0.27272727 0.30000000 0.30000000 0.30000000
## [67] 0.30769231 0.32608696 0.31914894 0.37500000 0.28571429
```