Lab2

Bingying Liu

1/27/2020

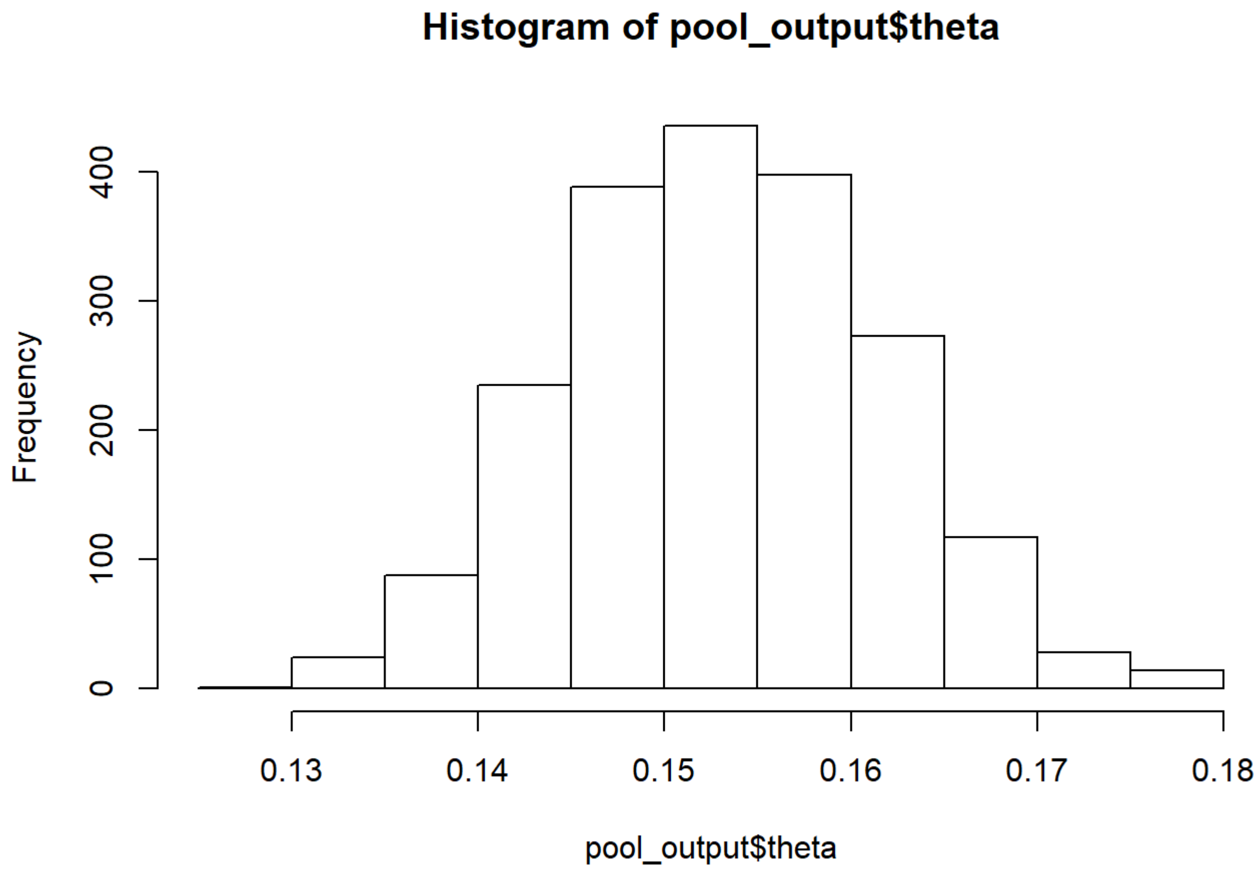
Exericise 1: Plot a histogram of θ from the rstan object called pool_output. Describe the distribution.

```
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)

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

```
## [1] 0.1531599
```

```
# histogram
hist(pool_output$theta)
```



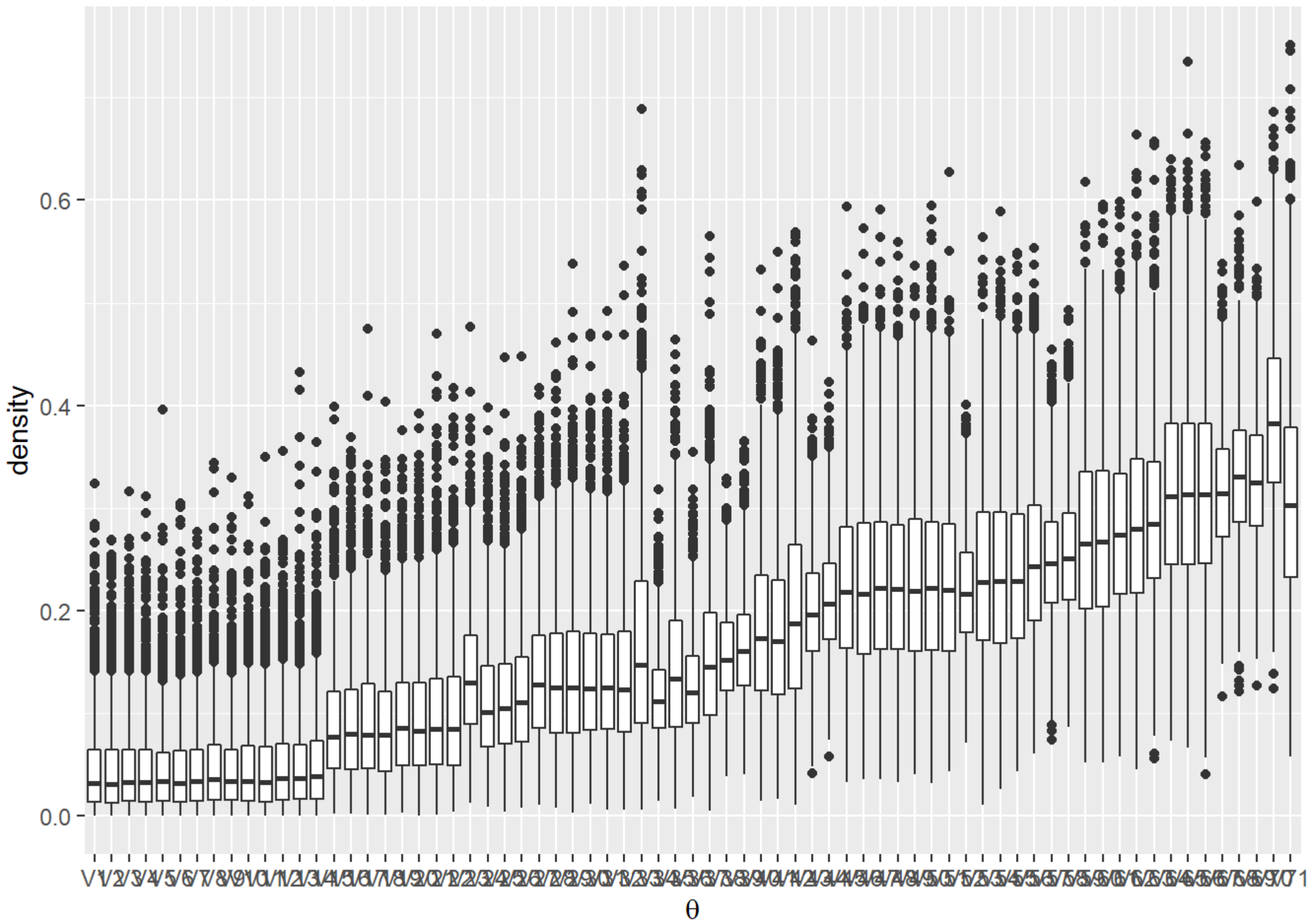
The distribution of pooled theta is approximately normal and centered around posterior mean which is 0.1536.

Exercise 2: Visualize the posterior distribution of θi

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

```
##  [1] 0.04630633 0.04602766 0.04635534 0.04580140 0.04525329 0.04516015
##  [7] 0.04644470 0.04895981 0.04679344 0.04789964 0.04796668 0.05064168
## [13] 0.04985893 0.05325999 0.08987400 0.09153296 0.09260737 0.09013861
## [19] 0.09662727 0.09570216 0.09963700 0.09923562 0.13727547 0.11150725
## [25] 0.11416895 0.11937775 0.13652992 0.13568908 0.13596845 0.13546465
## [31] 0.13615713 0.13585466 0.16839706 0.11674151 0.14537608 0.12565786
## [37] 0.15556345 0.15692535 0.16366602 0.18338971 0.18051954 0.20140723
## [43] 0.19957209 0.21102413 0.22628426 0.22709550 0.22812676 0.22794031
## [49] 0.22906328 0.22959928 0.22641659 0.21932524 0.23618122 0.23578901
## [55] 0.23768171 0.25048842 0.24913499 0.25475021 0.27238737 0.27293431
## [61] 0.27845591 0.28609394 0.29183265 0.31813294 0.31857597 0.31804119
## [67] 0.31616127 0.33310544 0.32723132 0.38548284 0.31061690
```

```
# boxplot of theta_i
nopool_df <- stack(as.data.frame(nopool_output$theta))
ggplot(nopool_df) +
  geom_boxplot(aes(x = ind, y = values))+
  ylab('density')+
  xlab(expression(theta))
```



Different posterior distributions of θ_i are plotted here. Since different group of rats have different probabilities of developing tumor, the sampling distributions for each group are different and give noninformative prior, the posterior distribution comes entirely from data. Each point in the boxplot represents outliers, which are 1.5*IQR (quantile) away from the box.

Exercise 3: How are the two stan files different?

Since in unpooled data, we assume different groups of rats have different probabilities of developing tumor while we assume same probabilities in pooled data. Below is what is shown in two stan files.

In nopool stan, parameters { vector<lower=0, upper=1>[n] theta; // chance of success (unpooled) }

In pool stan, parameters { real<lower=0, upper=1> theta; // chance of success (pooled) }

Exercise 4: What observable quantity do the parameter a and b represent about our prior beliefs?

‘a’ represents the number of rats that have developed cancer in prior samples while ‘b’ represents the number of rats that haven’t developed cancer in prior samples.

Exercise 5: What do we actually observe in the rat tumor data with respect to these quantities?

We observe in the posterior there are ‘a + sum of (yi)’ rats that have developed cancer and ‘b + n - sum of (yi)’ rats that haven’t developed cancer.

Exercise 6: How well do our different prior beliefs – the ones represented by the different parameter settings above – match up with the data?

Posterior distribution will center around prior mean a/(a+b) if dataset is small (prior has a strong influence to posterior) whereas if dataset is large, posterior distribution will center around sample mean.

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 θ and θi and comparing these values to the values obtained with the Bayesian approach:

We observed such a difference between pool and nopool approach because their underlying sampling distributions are different.

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

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

```
# approach 2
mle.2 <- y/N #
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
```

The posterior mean values of bayesian approach and MLE are the same because of the noninformative prior. MLE only takes into account data not prior, but since prior is a uniform distribution, so the results for bayesian and MLE approach are the same.