STATS 451 Homework 3

Yuzhou Peng

2024-02-13

Problem 1

$$\theta \text{ is the proportion of girls in total births.}$$
$$\text{Let } y_i = \begin{cases} 1, & \text{i-th birth is a girl} \\ 0, & \text{otherwise} \end{cases}$$
$$y_i \overset{\text{iid}}{\sim} Bernoul(\theta)$$
$$\text{The likelihood is given by: } \mathbb{P}(y_1, y_2 \ldots y_{1000}|\theta) = \prod_{i=1}^{1000} \mathbb{P}(y_i|\theta) = \theta^{450}(1-\theta)^{550}$$
$$\text{The posterior is given by: } \mathbb{P}(\theta|y_1, y_1 \ldots y_{1000}) = \frac{\mathbb{P}(y_1, y_2 \ldots y_{1000}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(y)} = b * \theta^{450}(1-\theta)^{550} \text{ where b is the normalizing constant.}$$
$$\theta|y_1, y_2 \ldots y_{1000} \sim Beta(451, 551)$$

Problem 2

```
# According to we have derived in problem 1, the posterior follows distribution of Beta(451, 551)
s1 <- 451
s2 <- 551

expect_pos <- s1/(s1 + s2)
variance_pos <- s1*s2/((s1 + s2)^2 * (s1 + s2 + 1))

expect_pos
```

```
## [1] 0.4500998
```

```
variance_pos
```

```
## [1] 0.0002467697
```

```
p_ci <- c(0.025, 0.975)
q_ci_pos <- qbeta(p_ci, shape1 = s1, shape2 = s2)

q_ci_pos
```

```
## [1] 0.4194118 0.4809764
```

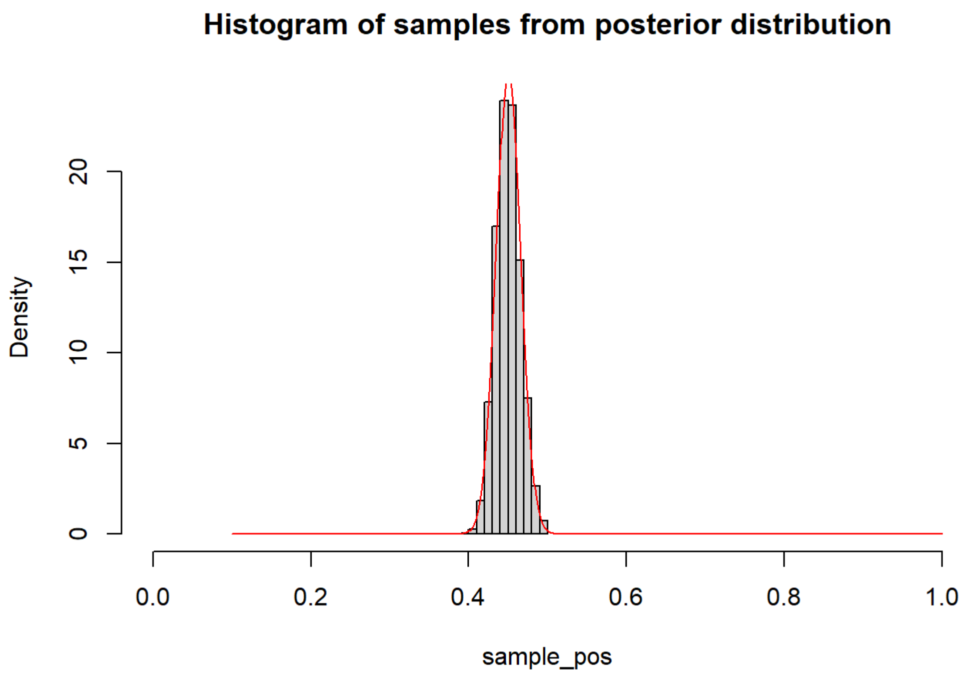
The expectation and variance of posterior distribution are 0.450 and 0.000247
A 95% confidence interval is [0.419, 0.481]

Problem 3

```
set.seed(451)

sample_pos <- rbeta(2000, shape1 = s1, shape2 = s2)
sample_step <- seq(0.1,1, by = 0.001)

hist(sample_pos,
     freq = F,
     xlim = c(0,1),
     main = "Histogram of samples from posterior distribution")
lines(sample_step, dnorm(sample_step, mean = expect_pos, sd = (variance_pos)^(1/2)), col = "red")
```



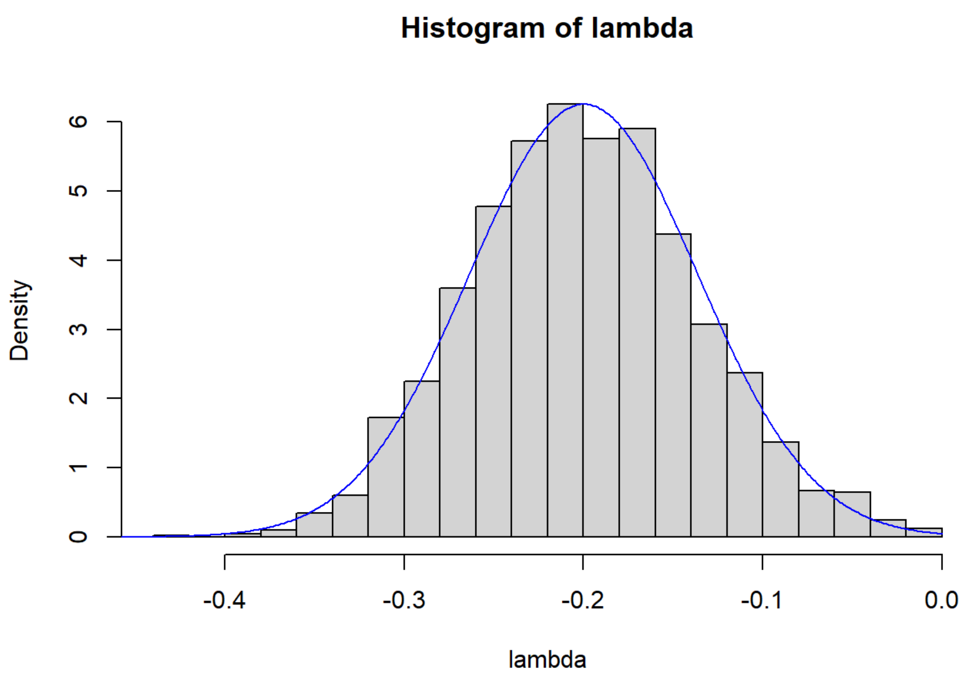
The histogram looks very similar to Gaussian distribution

Problem 4

```
lambda <- log(sample_pos/(1-sample_pos))
lambda_step <- seq(-0.5, 0, by = 0.001)

mean_lambda <- mean(lambda)
variance_lambda <- var(lambda)
ci_lambda <- quantile(lambda, probs = c(0.02, 0.98))

hist(lambda,
     freq = F,
     breaks = 20)
lines(lambda_step, dnorm(lambda_step, mean = mean_lambda, sd = (variance_lambda^(1/2))), col = "blue")
```




```
mean_lambda
```

```
## [1] -0.1999637
```

```
variance_lambda
```

```
## [1] 0.004070059
```

```
ci_lambda
```

```
##          2%         98%
## -0.32552868 -0.05855922
```

The histogram is similar to Gaussian distribution
Expected value is -0.200
Variance is 0.00407
A 96% confidence interval is [-0.326, -0.0586]

Problem 5

```
pbeta(0.485, s1, s2)
```

```
## [1] 0.9866025
```

Based on the data, we have 98.66% confidence to claim that proportion of girl births under the special condition is less than 0.485

Problem 6

6.1

```
s1_prime <- 5
s2_prime <- 7

expect_pos_prime <- s1_prime/(s1_prime + s2_prime)
variance_pos_prime <- s1_prime*s2_prime/((s1_prime + s2_prime)^2 * (s1_prime + s2_prime + 1))

expect_pos_prime
```

```
## [1] 0.4166667
```

```
variance_pos_prime
```

```
## [1] 0.01869658
```

```
qbeta(p_ci, shape1 = s1_prime, shape2 = s2_prime)
```

```
## [1] 0.1674881 0.6920953
```

The expectation and variance of posterior distribution are 0.417 and 0.0187
A 95% confidence interval is [0.167, 0.692]

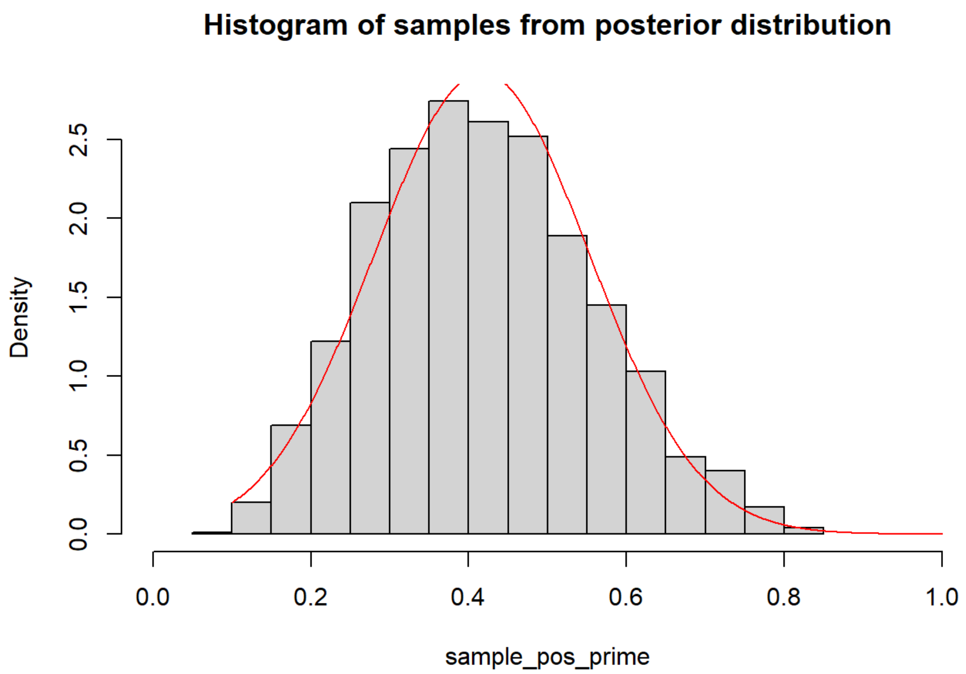
In comparison with posterior of larger sample, variance is much larger and confidence interval is much wider.
The expected value is more affected by the prior mean.

6.2

```
set.seed(451)

sample_pos_prime <- rbeta(2000, shape1 = s1_prime, shape2 = s2_prime)
sample_step <- seq(0.1,1, by = 0.001)

hist(sample_pos_prime,
     freq = F,
     xlim = c(0,1),
     main = "Histogram of samples from posterior distribution")
lines(sample_step, dnorm(sample_step, mean = expect_pos_prime, sd = (variance_pos_prime)^(1/2)), col = "red")
```

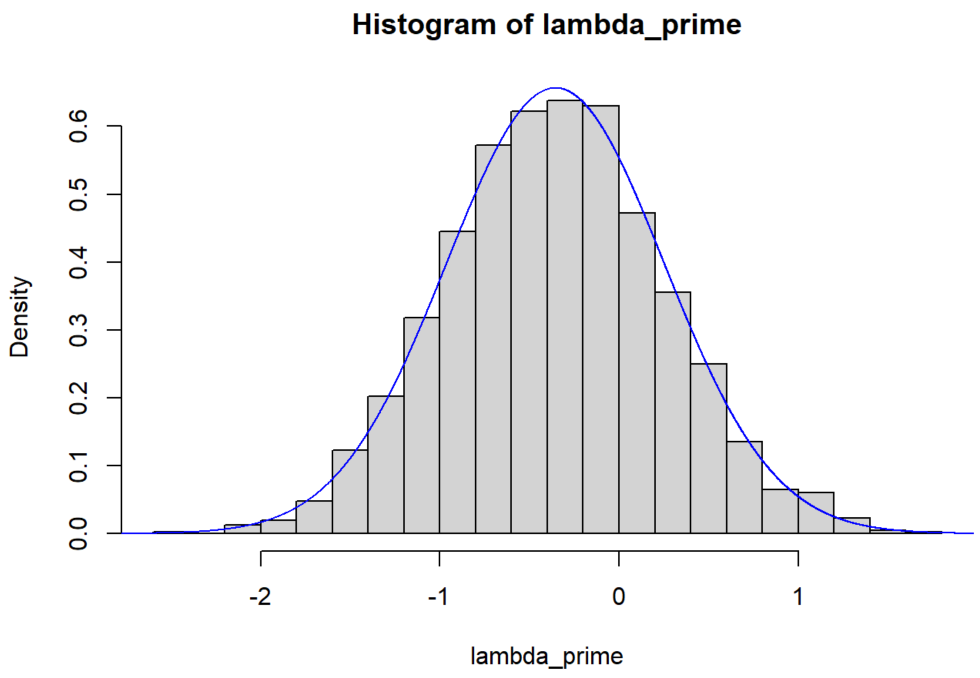


6.3

```
lambda_prime <- log(sample_pos_prime/(1-sample_pos_prime))
lambda_step_prime <- seq(-3, 3, by = 0.001)

mean_lambda_prime <- mean(lambda_prime)
variance_lambda_prime <- var(lambda_prime)
ci_lambda_prime <- quantile(lambda_prime, probs = c(0.02, 0.98))

hist(lambda_prime,
     freq = F,
     breaks = 20)
lines(lambda_step_prime, dnorm(lambda_step_prime, mean = mean_lambda_prime, sd = (variance_lambda_prime^(1/2))), col = "blue")
```




```
mean_lambda_prime
```

```
## [1] -0.3550896
```

```
variance_lambda_prime
```

```
## [1] 0.3691726
```

```
ci_lambda_prime
```

```
##         2%        98%
## -1.5587518  0.9739823
```

The histogram is similar to Gaussian distribution
Expected value is -0.355
Variance is 0.369
A 96% confidence interval is [-1.59, 0.974]

In comparison with posterior of larger sample, variance is much larger and confidence interval is much wider.

6.4

```
pbeta(0.485, s1_prime, s2_prime)
```

```
## [1] 0.6907652
```

Compared with the posterior of larger sample, we have lower confidence (69.08% compared with 98.66%) to claim the proportion of girl birth is lower than 0.485 even though the posterior mean is less than large sample posterior mean.