bayesian_stats_ch4_montecarlo_approximation

inoue jin

2022-12-03

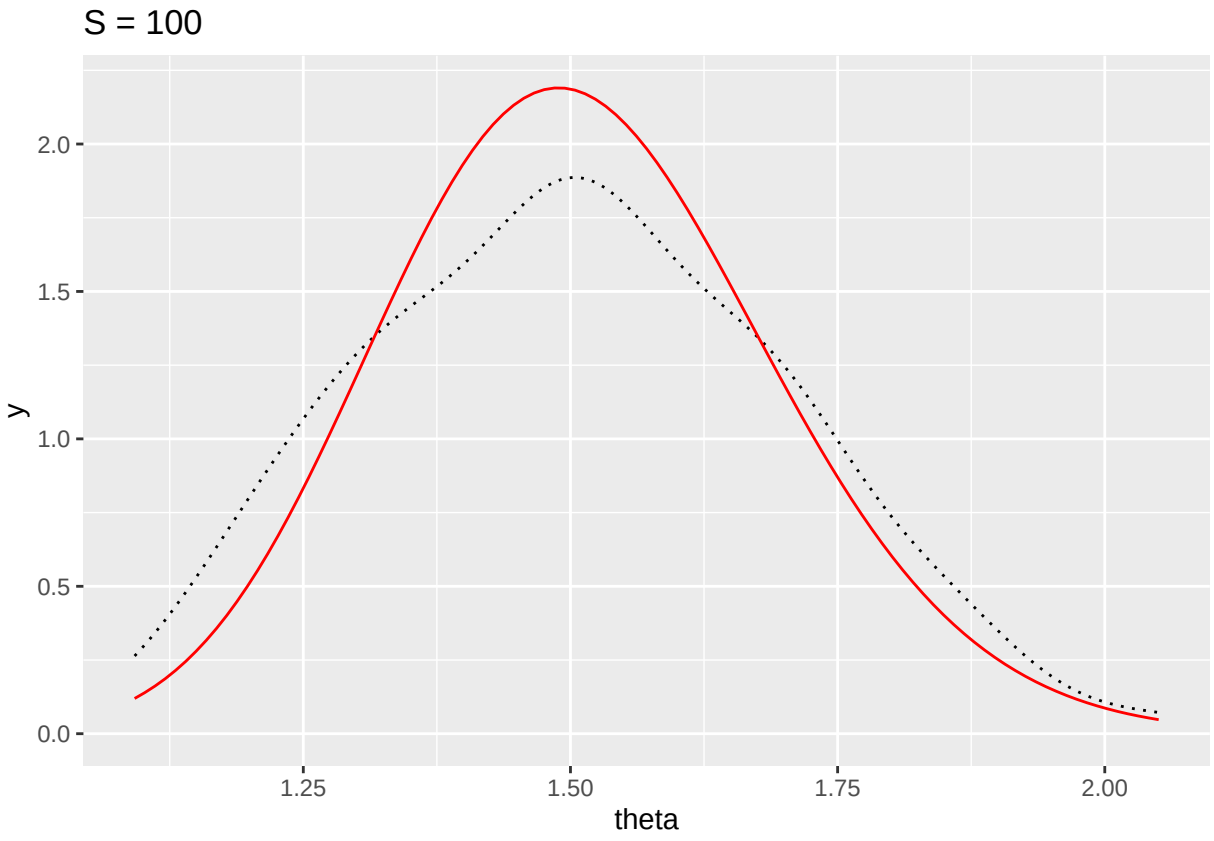
4.1 モンテカルロ法

```
a = 68
b = 45



s100 <- tibble(theta = rgamma(100, a, b))
s1000 <- tibble(theta = rgamma(1000, a, b))
s10000 <- tibble(theta = rgamma(10000, a, b))

s100
```

■ガンマ分布の真の密度関数と経験分布によるモンテカルロ近似

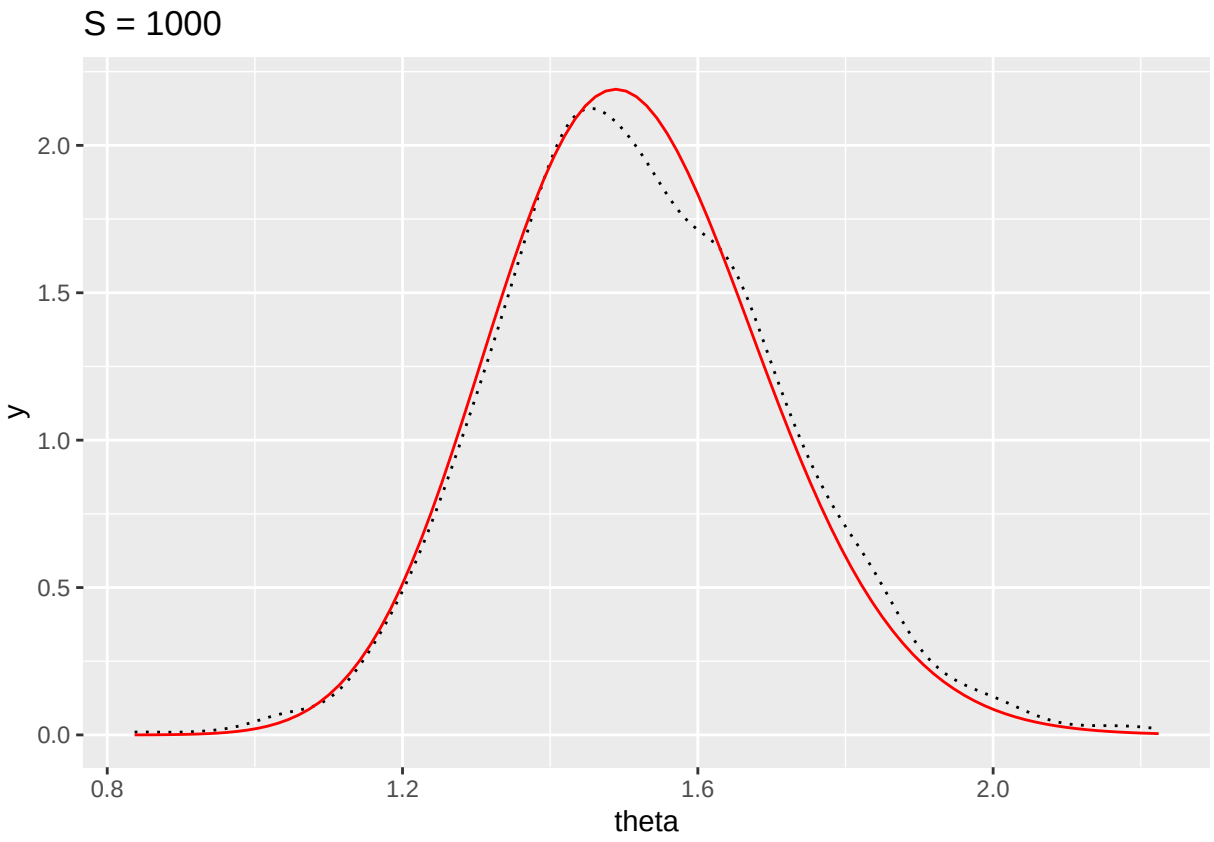
```
## # A tibble: 100 x 1
##    theta
##    <dbl>
##  1  1.49
##  2  1.64
##  3  1.88
##  4  1.49
##  5  1.35
##  6  1.69
##  7  1.69
##  8  1.23
##  9  1.46
## 10  1.48
## # ... with 90 more rows
```

```r
# curve(dgamma(x, a, b), -2, 4)

gg <- ggplot(data.frame(x = c(-2, 4)), aes(x = x)) +
  xlim(c(-2, 4)) +
  ylim(c(0, 3)) +
  stat_function(fun = dgamma, args = list(a, b))

gg100 <- s100 %>%
  ggplot(aes(x = theta)) +
  geom_density(linetype = "dotted")+
  stat_function(fun = dgamma, args = list(a, b), color = "red") +
  labs(title = "S = 100") +
  xlab("theta")
gg100
```




```r
gg1000 <- s1000 %>%
  ggplot(aes(x= theta)) +
  geom_density(linetype = "dotted") +
  stat_function(fun = dgamma, args = list(a, b), color = "red") +
  labs(title = "S = 1000")+
```

```
  xlab("theta")

gg1000
```



数値的な評価

まず、モンテカルロ法によって得られる近似値を、事後要約統計量の解析的に得られる解と比較する。

- モデル $Y_1, \ldots, Y_n \mid \theta \sim$ i.i.d.$Poisson(\theta)$
- θ の事前分布は $gamma(a, b)$
- この時、$Y_1 = y_1, \ldots, Y_n = y_n$ を観測した時の事後分布は、$gamma(a + \Sigma y_i, b + n)$ となる
- 事後平均は $(a + \Sigma y_i)/(b + n) = 1.51$ となる

```
a <- 2
b <- 1
sy <- 66
n <- 44

theta_mc10 <- rgamma(10, a + sy, b+n)
theta_mc100 <- rgamma(100, a+sy, b+n)
theta_mc1000 <- rgamma(1000, a+sy, b+n)
```

```r
mean(theta_mc10)
```

```
## [1] 1.469769
```

```r
mean(theta_mc100)
```

```
## [1] 1.482593
```

```r
mean(theta_mc1000)
```

```
## [1] 1.517637
```

```r
## theta < 1.75 の確率を計算
# Pro(theta < 1.75)



pgamma(q = 1.75, a+sy, b+n)
```

```
## [1] 0.8998286
```

```r
mean(theta_mc10 < 1.75)# \sum 1{theta_mc10 < 1.75}*(1/n)
```

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

```r
## 事後分布による 95% 信用領域
qgamma(c(0.025, 0.975), a+sy, b+n)
```

```
## [1] 1.173437 1.890836
```

```r
# monte carlo 標本から 95% 信用領域を求める

quantile(theta_mc10, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 1.152970 1.701792
```

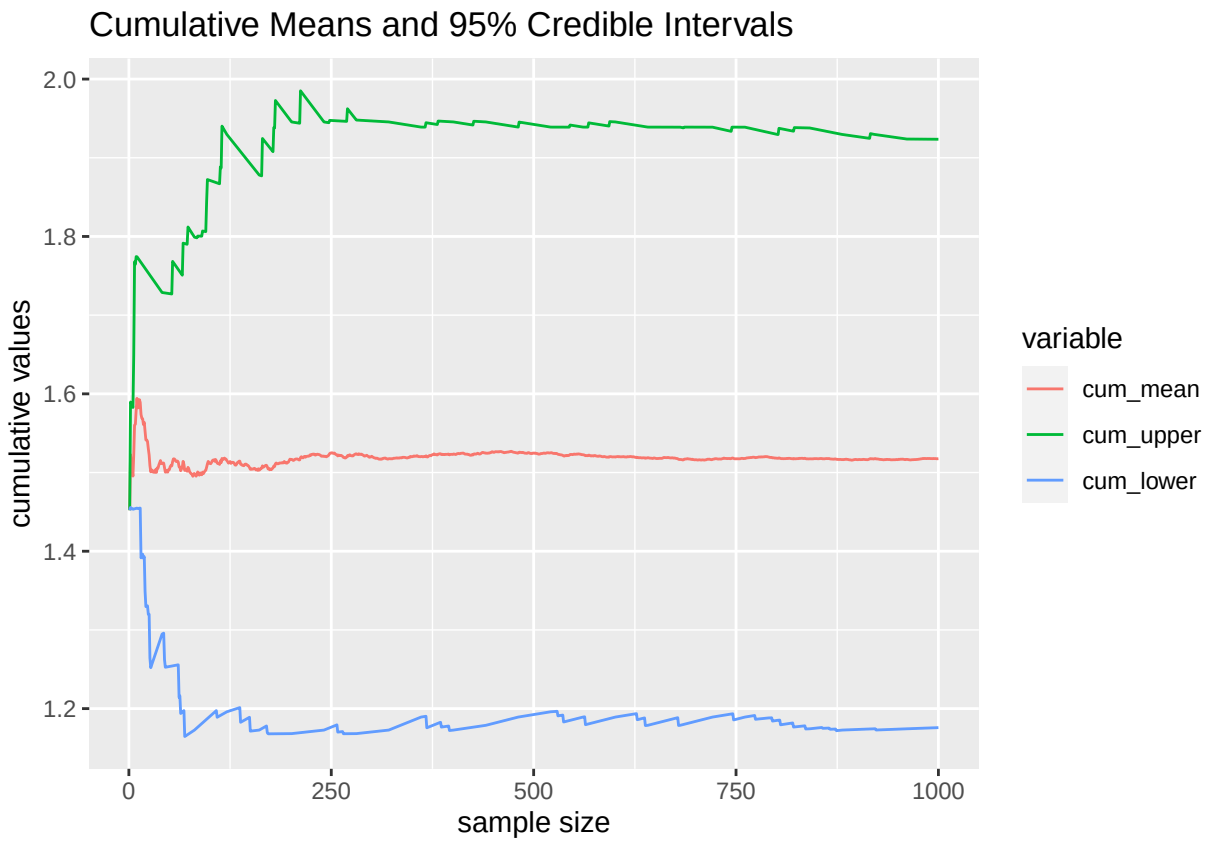
```r
# quantile(theta_mc100, 0.975) %>% as.double()
```

```r
# list で、ここの累積リストに対して quantile を適用して累積を計算する

# theta_mc1000
# theta_mc1000[1:1000]

theta_list <- list()# 1000 * 1000 のリスト作成
```

```r
# O(N^2)の計算時間
for(i in 1:1000){
  theta_list[[i]] <- theta_mc1000[1:i]
}



# theta_list %>% glimpse()
# theta_list[[1]]
# map(.x = theta_list, ~as.double(quantile(.x, 0.975)))

# 累積平均、累積97.5%分位点、累積2.5%分位点をまとめて作成

cum_df <- tibble(n = 1:1000,
                 cum_mean = cumsum(theta_mc1000)/(1:1000),
                 cum_upper = unlist(map(.x = theta_list, ~as.double(quantile(.x, 0.975)))),
                 cum_lower = unlist(map(.x = theta_list, ~as.double(quantile(.x, 0.025)))))

# cum_df %>% summary()

# {reshape2}でlong型にする

data_long <- reshape2::melt(cum_df, id = "n")

data_long %>% glimpse()
```

```
## Rows: 3,000
## Columns: 3
## $ n        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ variable <fct> cum_mean, cum_mean, cum_mean, cum_mean, cum_mean, cum_mean, c~
## $ value    <dbl> 1.452107, 1.522629, 1.511162, 1.503330, 1.495472, 1.522360, 1~
```

```r
gg_cumsum <- data_long %>%
  ggplot(aes(x = n, y = value, color = variable))+
  geom_line() +
  xlab("sample size")+
  ylab("cumulative values")+
  labs(title = "Cumulative Means and 95% Credible Intervals")

gg_cumsum
```



このような図は、モンテカルロ推定値が真の値に収束しているかを判定する上で役に立つ。また、モンテカルロ標準誤差を用いることで、事後平均の近似の精度を評価することができる。

- $\bar{\theta} = \sum_{s=1}^{S} \theta^{(s)}/S$ をモンテカルロ標本の平均とすると、中心極限定理より、$\bar{\theta}$ は期待値 $\mathbb{E}[\theta \mid y_1, \dots, y_n]$、標準偏差 $\sqrt{Var[\theta \mid y_1, \dots, y_n]/S}$ で近似でき、モンテカルロ標準偏差はこの標準偏差を標本で近似したものである。
- $\hat{\sigma}^2 = \Sigma(\theta^{(s)} - \bar{\theta})^2/(S-1)$ を $Var[\theta \mid y_1, \dots, y_n]$ のモンテカルロ推定値とすると、モンテカルロ標準誤差は $\sqrt{\hat{\sigma}^2/S}$ となる。
- θ の事後平均の 95% 近似モンテカルロ信頼区間は $\bar{\theta} \pm 2\sqrt{\hat{\sigma}^2/S}$ である。
- モンテカルロ標本のサイズの基準としては、このモンテカルロ標準誤差が $\mathbb{E}[\theta \mid y_1, \dots, y_n]$ を報告する制度よりも小さくなるように、S を十分に大きく取るというようなことが考えられる。

4.2 任意の関数に対する事後推測

二項モデルを用いるとき、次のような対数オッズに関心があることがある。

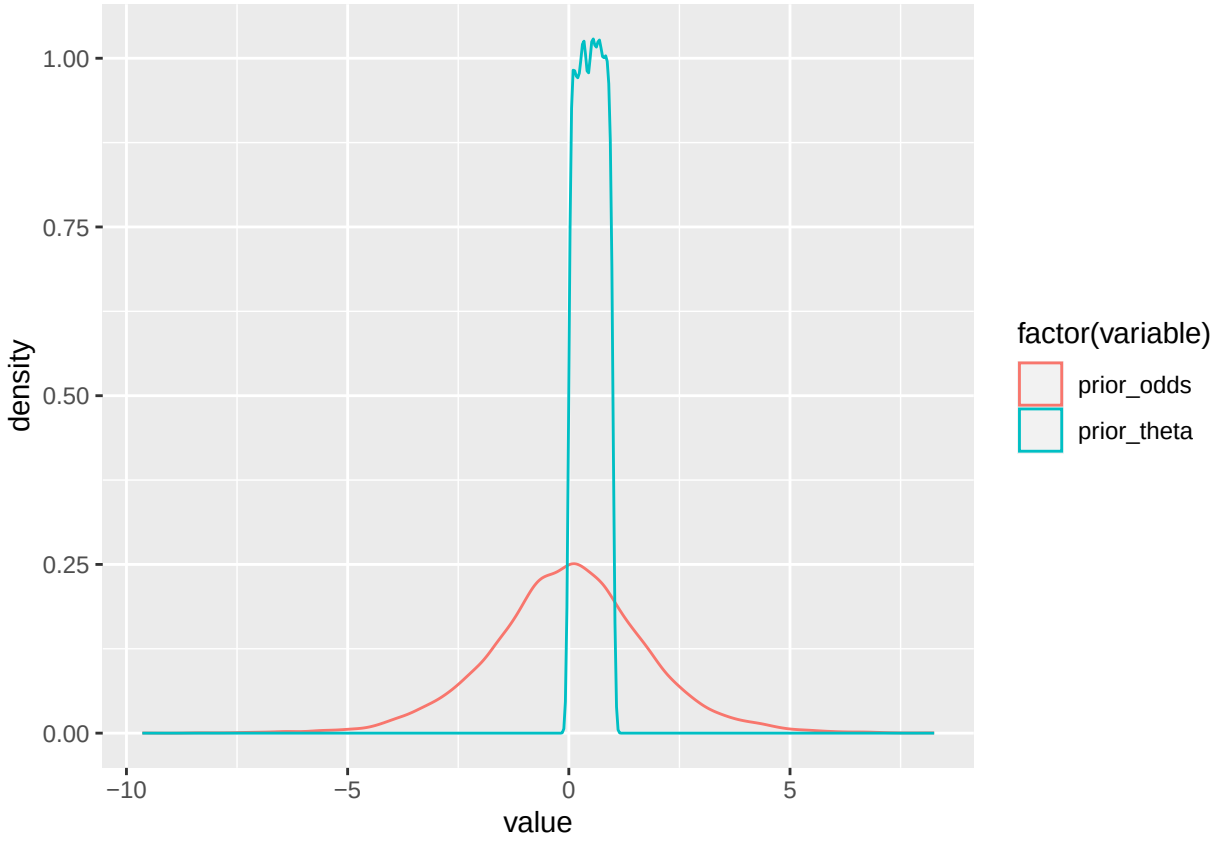
$$\log \text{odds}(\theta) = \log \frac{\theta}{1-\theta} = \gamma$$

■例：対数オッズ　1998 年の GSS(General Social Survey) の回答者の 54% は、宗教的志向がプロスタントであると回答し、非プロテスタントは少数派にとどまった。調査では、公立学校で経典を読むことを州または地方政府が要求することを禁止する最高裁判所の判決に同意するかどうかが尋ねられ、宗教的少数派（非プロテスタント）の $n = 860$ 人のうち、$y = 441$ が同意したのに対し、プロテスタントでは 1011 人のうち、353 人が判決に同意した。

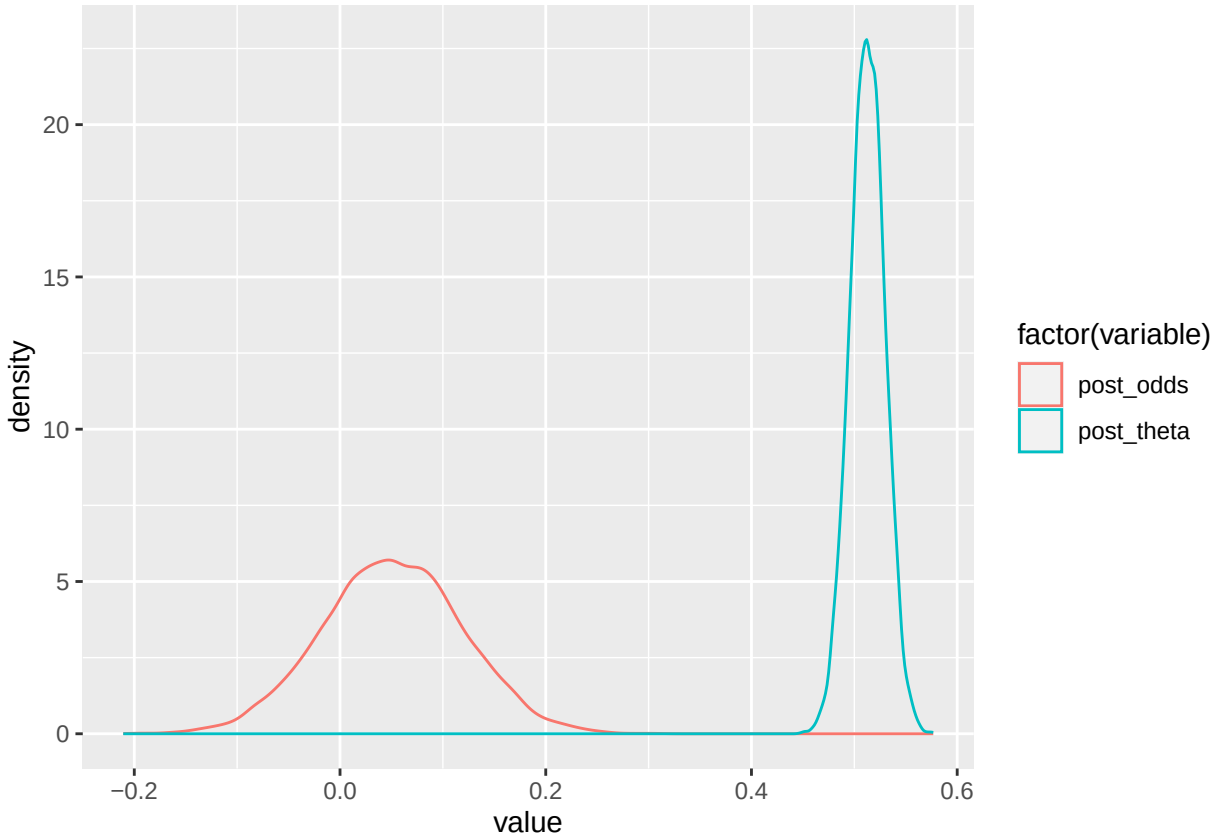
少数派の母集団に対応する母比率を θ とする。標本モデルに二項分布を、事前分布に一様分布を用いると、θ の事後分布は、$\text{beta}(y+1 = 442, n-y+1 = 220)$ となる。

```
a <- 1
b <- 1

theta_prior_mc <- rbeta(10000, a, b)#Beta(a,b) で a=b=1 とすると Uniform(0,1) となる
gamma_prior_mc <- log(theta_prior_mc/(1-theta_prior_mc))

n <- 860
y <- 441

theta_post_mc <- rbeta(10000, y+1, n-y+1)
gamma_post_mc <- log(theta_post_mc/(1-theta_post_mc))

df_prior <- tibble(prior_odds = gamma_prior_mc,
             prior_theta = theta_prior_mc) %>%
  mutate(id = row_number())

df_post <- tibble(post_odds = gamma_post_mc,
                  post_theta = theta_post_mc) %>%
  mutate(id = row_number())

prior_long <- df_prior %>%
  reshape2::melt(id = "id")

post_long <- df_post %>%
  reshape2::melt(id = "id")

# prior_long

prior_long %>%
  ggplot(aes(x = value)) +
```

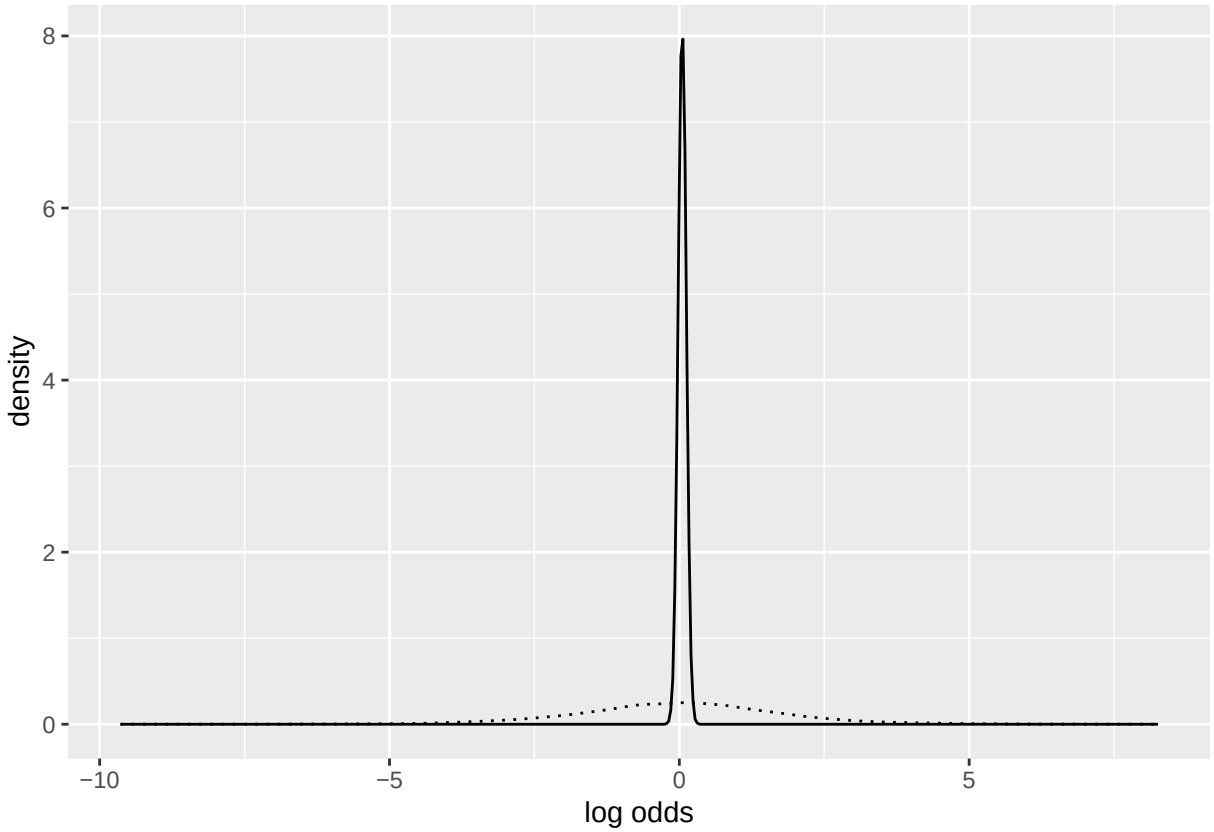
```
  geom_density(aes(color = factor(variable)))
```




```
post_long %>%
  ggplot(aes(x = value)) +
  geom_density(aes(color = factor(variable)))
```




```
odds <- tibble(prior_odds = gamma_prior_mc,
               post_odds = gamma_post_mc)

odds %>%
  ggplot()+
  geom_density(aes(x = prior_odds),linetype = "dotted") +
  geom_density(aes(x = post_odds))+
  xlab("log odds")
```


density
8
6
4
2
0
−10
−5
0
5
log odds