Bayesian Assignment 1

Gavin Connolly

12/02/2022

Introduction

A study was carried out to measure how successful a new drug is to lower hypertension in comparison to an existing drug. The new drug was administered to 300 patients and of these 145 recorded a reduction in hypertension. While the existing drug was administered to a total of 350 patients and 144 of these recorded a reduction in hypertension. Our key question in interest is to assess whether the new drug is more successful in reducing hypertension than the existing drug. To do this we will model the outcomes for each drug using independent binomial distributions. Initially, we will use a uniform prior distribution for the success probability (in the binomial distribution) for each drug. Again, we will assume that these are independent of each other.

Question 1

Use Stan to estimate the posterior distribution for the success probability for each group. Provide appropriate plots and summaries. Comment briefly on these.

Create data list based on experiment data:

```
d_bin <- list(N1 = 300, y1 = 145, N2 = 350, y2 = 144)
```

We write the following stan code to model our problem.

```
writeLines(readLines("binomHypertension1.stan"))
```

```
## //  Comparison of two groups with Binomial
## data {
##   int<lower=0> N1;
##   int<lower=0> y1;
##   int<lower=0> N2;
##   int<lower=0> y2;
## }
## parameters {
##   real<lower=0,upper=1> theta1;
##   real<lower=0,upper=1> theta2;
## }
## model {
##   theta1 ~ beta(1,1);
##   theta2 ~ beta(1,1);
##   y1 ~ binomial(N1,theta1);
##   y2 ~ binomial(N2,theta2);
## }
## generated quantities {
##   real oddsratio;
##   real difference;
##
##   oddsratio = (theta2/(1-theta2))/(theta1/(1-theta1));
##   difference = theta2-theta1;
## }
```

```
fit_bin <- stan(file = 'binomHypertension1.stan', data = d_bin, iter = 10000, seed = SEED)

monitor(fit_bin, probs = c(0.1, 0.5, 0.9))
```

```
## Inference for the input samples (4 chains: each with iter = 10000; warmup = 0):
##
##                 Q5    Q50    Q95   Mean  SD  Rhat Bulk_ESS Tail_ESS
## theta1         0.4    0.5    0.5    0.5 0.0     1    18284    13542
## theta2         0.4    0.4    0.5    0.4 0.0     1    18175    13586
## oddsratio      0.6    0.7    1.0    0.8 0.1     1    17801    13340
## difference    -0.1   -0.1    0.0   -0.1 0.0     1    17806    13428
## lp__        -450.7 -448.4 -447.7 -448.7 1.0     1     8588    11003
##
## For each parameter, Bulk_ESS and Tail_ESS are crude measures of
## effective sample size for bulk and tail quantities respectively (an ESS > 100
## per chain is considered good), and Rhat is the potential scale reduction
## factor on rank normalized split chains (at convergence, Rhat <= 1.05).
```

We calculate the mean & variance of the distribution for the true proportion of successful trials for the new drug.

```
draws <- as.data.frame(fit_bin)

mean(draws$theta1)
```

```
## [1] 0.4836122
```

```
var(draws$theta1)
```

```
## [1] 0.0008314647
```

We calculate the 95% Highest Posterior Density Interval for the mean proportion of successes for the new drug.

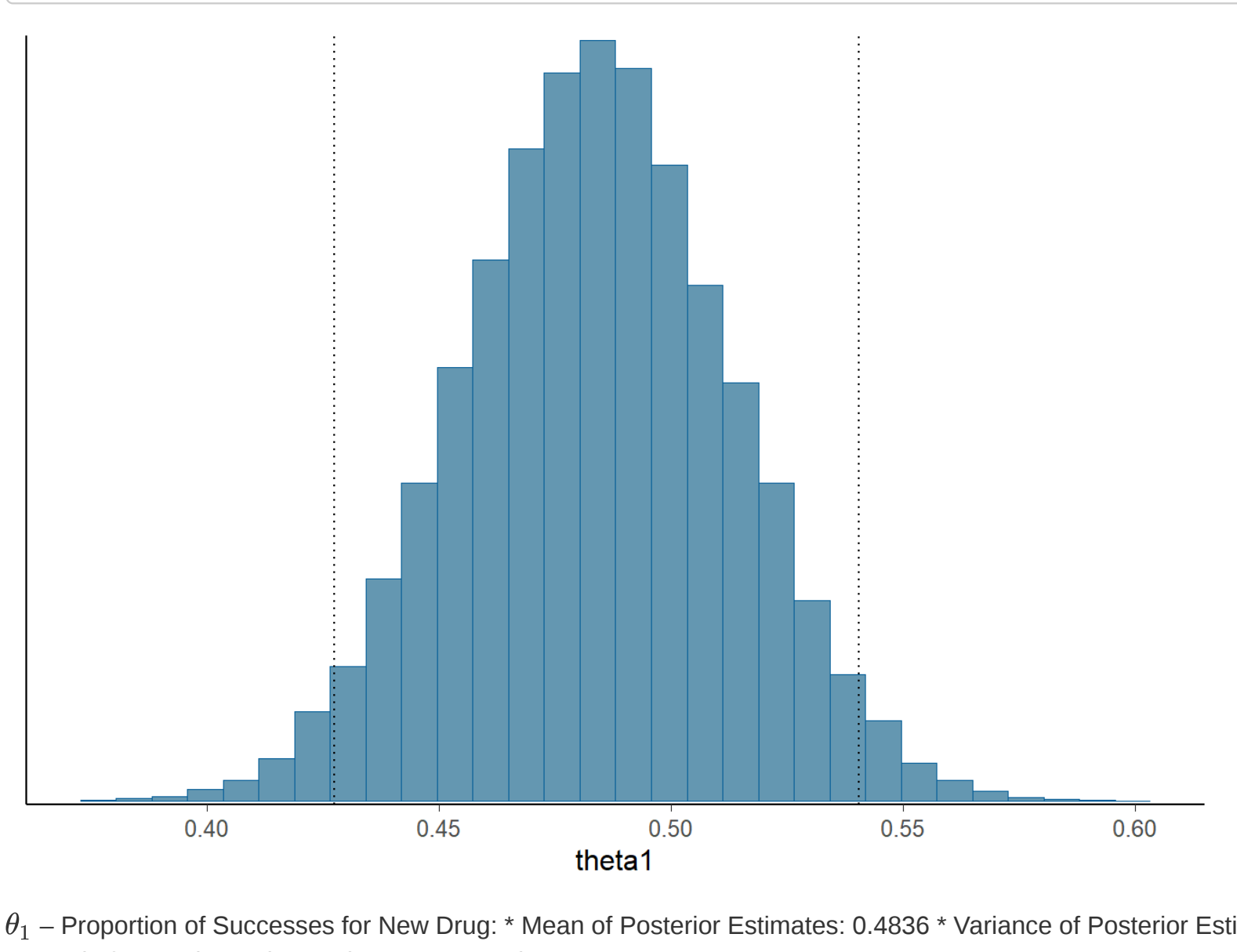
```
# Calculate 95% Credible Interval for Posterior Estimate of theta of New Data
CredInts = HDInterval::hdi(draws, 0.95)
CredIntNew = CredInts[,1]
CredIntNew
```

```
##     lower     upper
## 0.4273970 0.5404568
```

We plot the posterior distribution for the true proportion of successes for the new drug, along with the 95% highest density credible interval.

```
mcmc_hist(draws, pars = 'theta1') +
  geom_vline(xintercept = CredIntNew[1], linetype='dotted') + # plot the 95% credible interval
  geom_vline(xintercept = CredIntNew[2], linetype='dotted')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



θ_1 – Proportion of Successes for New Drug: * Mean of Posterior Estimates: 0.4836 * Variance of Posterior Estimates: 0.000831 * 95% Credible Interval of Posterior Estimate: (0.4274, 0.5405)

```
mean(draws$theta2)
```

```
## [1] 0.4119253
```

```
var(draws$theta2)
```

```
## [1] 0.0006793381
```

We calculate the 95% Highest Posterior Density Interval for the mean proportion of successes for the old drug.

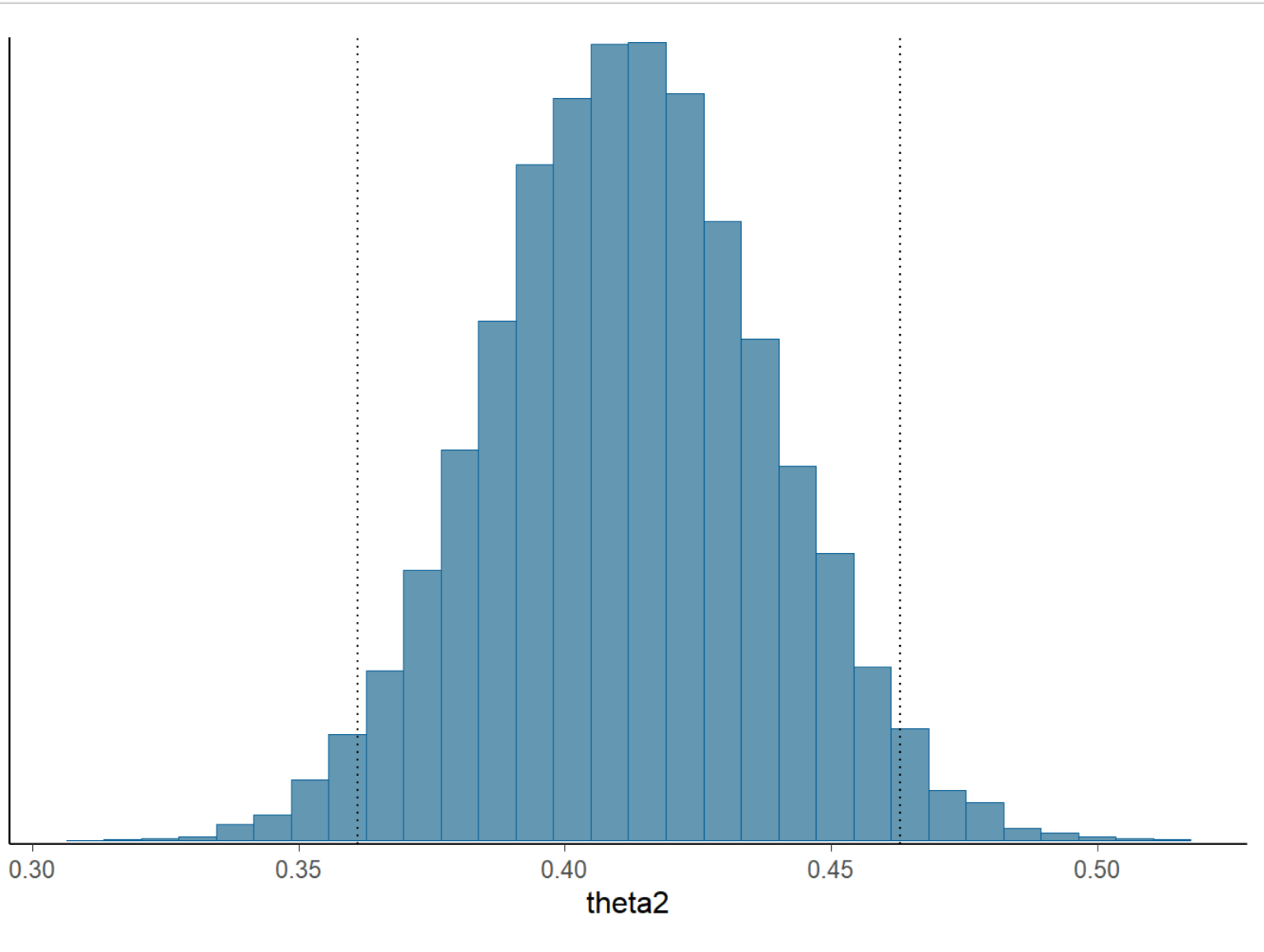
```
# Calculate 95% Credible Interval for Posterior Estimate of theta of Old Data
CredIntOld = CredInts[,2]
CredIntOld
```

```
##     lower     upper
## 0.3610456 0.4629518
```

We plot the posterior distribution for the true proportion of successes for the old drug, along with the 95% highest density credible interval.

```
mcmc_hist(draws, pars = 'theta2') +
  geom_vline(xintercept = CredIntOld[1], linetype='dotted') + # plot the 95% credible interval
  geom_vline(xintercept = CredIntOld[2], linetype='dotted')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



θ_2 – Proportion of Successes for Existing Drug: * Mean of Posterior Estimates: 0.4119 * Variance of Posterior Estimates: 0.000679 * 95% Credible Interval of Posterior Estimate: (0.361, 0.463)

We see from the posterior summary statistics that the mean proportion of successes for the new drug lies outside of the 95% credible interval for the proportion of successes for the existing drug, indicating that there is a significant difference in the distribution of posterior means. As the mean proportion of successes for the new drug is higher than that of the existing drug, this would imply that the new drug was significantly more effective at relieving hypertension in the sampled patients.

Question 2

Is there evidence to suggest that the new drug is more effective than the existing drug at reducing hypertension? To answer this, estimate the posterior odds ratio between the new drug and the old drug. Provide an appropriate plot and summary of this and comment briefly on your finding.

The Odds Ratio is given by the formula:

$$\frac{\theta_2}{1-\theta_2} \div \frac{\theta_1}{1-\theta_1}$$

Calculate the mean & variance of the odds ratio.

```
mean(draws$oddsratio)
```

```
## [1] 0.7566881
```

```
var(draws$oddsratio)
```

```
## [1] 0.01442796
```

Calculate the 95% Credible Interval.

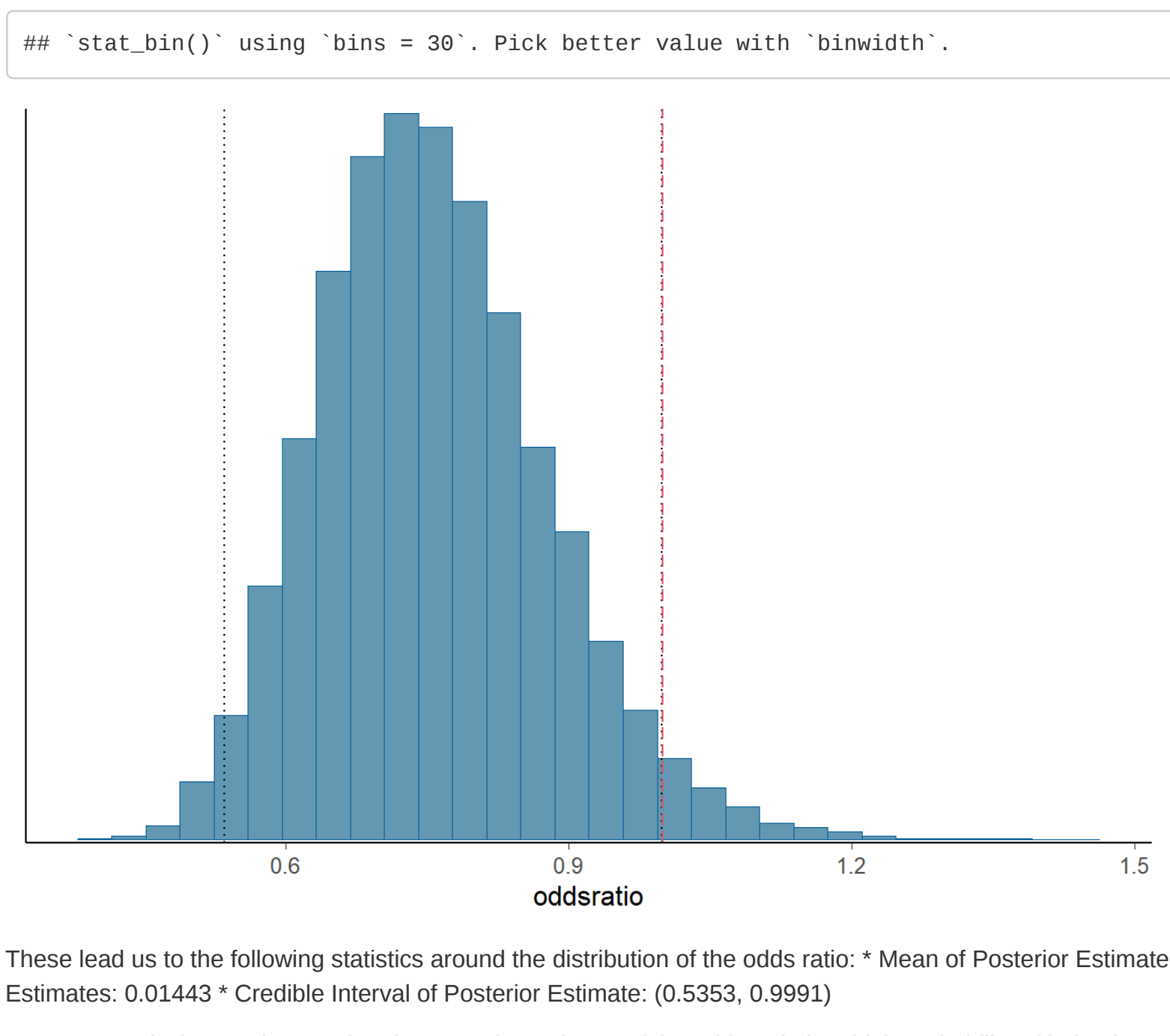
```
# Calculate 95% Credible Interval for Posterior Estimate of Odds Ratio
CredIntOdds = CredInts[,3]
CredIntOdds
```

```
##     lower     upper
## 0.5353354 0.9991042
```

Plot the posterior estimate of the odds ratio, along with the 95% CI.

```
mcmc_hist(draws, pars = 'oddsratio') +
  geom_vline(xintercept = CredIntOdds[1], linetype='dotted') + # plot the 95% credible interval
  geom_vline(xintercept = CredIntOdds[2], linetype='dotted') +
  geom_vline(xintercept = 1, linetype='longdash', col = '2') # plot x = 1 line (odds ratio of thetas equal)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



These lead us to the following statistics around the distribution of the odds ratio: * Mean of Posterior Estimates: 0.7567 * Variance of Posterior Estimates: 0.01443 * Credible Interval of Posterior Estimate: (0.5353, 0.9991)

From our analysis, we observe that the posterior estimate of the odds ratio has high probability of being less than 1 with an estimated posterior mean of 0.7566, which would indicate that the new drug was more effective at relieving hypertension symptoms in patients. We also notice that the value 1 lies just outside of our 95% credible interval for the uncertainty surrounding the posterior estimate for the odds ratio, meaning we can say the parameter odds ratio will lie below 1 with 95% probability.

Question 3

Similar to the previous question, now provide a posterior distribution of the difference between the success probabilities for the new and old drugs. Again, provide appropriate plots and summary statistics. Present a brief commentary.

Here we examine the difference in success probabilities given by: $\theta_2 - \theta_1$

Conducting similar analysis to before, we get the following results.

```
mean(draws$difference)
```

```
## [1] -0.0716869
```

```
var(draws$difference)
```

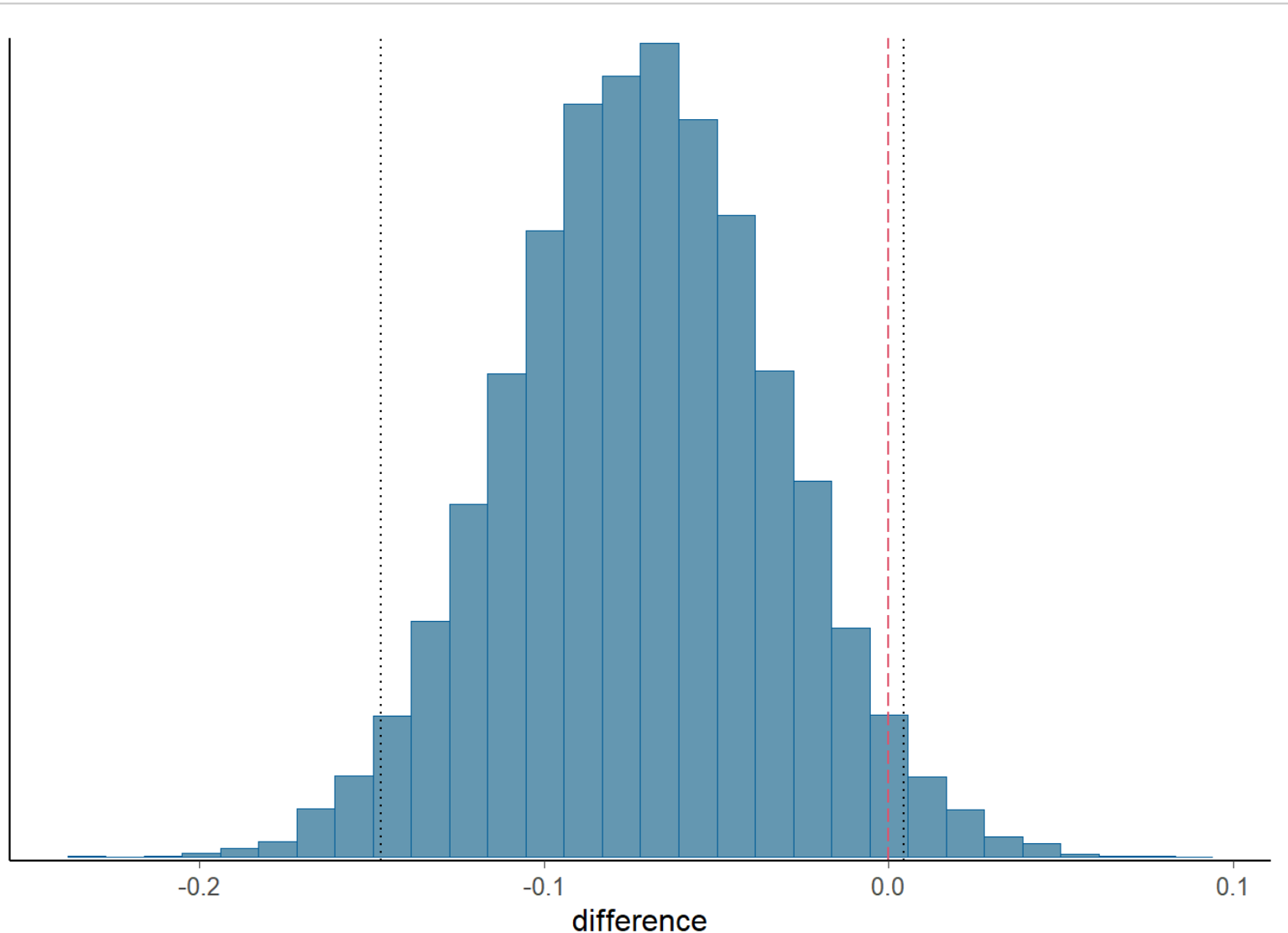
```
## [1] 0.001499622
```

```
CredIntDiff = CredInts[,4]
CredIntDiff
```

```
##        lower        upper
## -0.147287282  0.004454676
```

```
mcmc_hist(draws, pars = 'difference') +
  geom_vline(xintercept = CredIntDiff[1], linetype='dotted') + # plot the 95% credible interval
  geom_vline(xintercept = CredIntDiff[2], linetype='dotted') +
  geom_vline(xintercept = 0, linetype='longdash', col = '2') # plot x = 0 line (difference between thetas equals
 0)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- Mean of Posterior Estimates: -0.0717
- Variance of Posterior Estimates: 0.0015
- Credible Interval of Posterior Estimate: (-0.1473, 0.0045)

Once again from the posterior distribution of the difference $\theta_2 - \theta_1$, we see that the estimate of the posterior mean difference between θ_1 & θ_2 lies below 0 with high probability. This would again indicate that the new drug was more effective at treating hypertension symptoms in patients. However, we also observe the value 0 lies within our 95% credible interval for the difference which means we cannot say the difference $\theta_2 - \theta_1$ will lie below 0 with 95% probability.

Question 4

Let's now use a more informative beta prior. To do this, we will use the following interpretation of a beta(a, b) prior: a = no. of prior successes and b = no. of prior failures. Then we can interpret a + b as the total no. of prior trials. Following this interpretation, set a = 10 and b = 10. Repeat the analysis carried out in 2. and 3 again providing appropriate plots, summaries and commentary.

```
writeLines(readLines("binomHypertension2.stan"))
```

```
## //  Comparison of two groups with Binomial
## data {
##   int<lower=0> N1;
##   int<lower=0> y1;
##   int<lower=0> N2;
##   int<lower=0> y2;
## }
## parameters {
##   real<lower=0,upper=1> theta1;
##   real<lower=0,upper=1> theta2;
## }
## model {
##   theta1 ~ beta(10,10);
##   theta2 ~ beta(10,10);
##   y1 ~ binomial(N1,theta1);
##   y2 ~ binomial(N2,theta2);
## }
## generated quantities {
##   real oddsratio;
##   real difference;
##
##   oddsratio = (theta2/(1-theta2))/(theta1/(1-theta1));
##   difference = theta2-theta1;
## }
```

```
fit_bin2 <- stan(file = 'binomHypertension2.stan', data = d_bin, iter = 10000, seed = SEED)

monitor(fit_bin, probs = c(0.1, 0.5, 0.9))
```

```
## Inference for the input samples (4 chains: each with iter = 10000; warmup = 0):
##
##                 Q5    Q50    Q95   Mean  SD  Rhat Bulk_ESS Tail_ESS
## theta1         0.4    0.5    0.5    0.5 0.0     1    18284    13542
## theta2         0.4    0.4    0.5    0.4 0.0     1    18175    13586
## oddsratio      0.6    0.7    1.0    0.8 0.1     1    17801    13340
## difference    -0.1   -0.1    0.0   -0.1 0.0     1    17806    13428
## lp__        -450.7 -448.4 -447.7 -448.7 1.0     1     8588    11003
##
## For each parameter, Bulk_ESS and Tail_ESS are crude measures of
## effective sample size for bulk and tail quantities respectively (an ESS > 100
## per chain is considered good), and Rhat is the potential scale reduction
## factor on rank normalized split chains (at convergence, Rhat <= 1.05).
```

```
draws2 <- as.data.frame(fit_bin2)
mean(draws2$oddsratio)
```

```
## [1] 0.7688621
```

```
var(draws2$oddsratio)
```

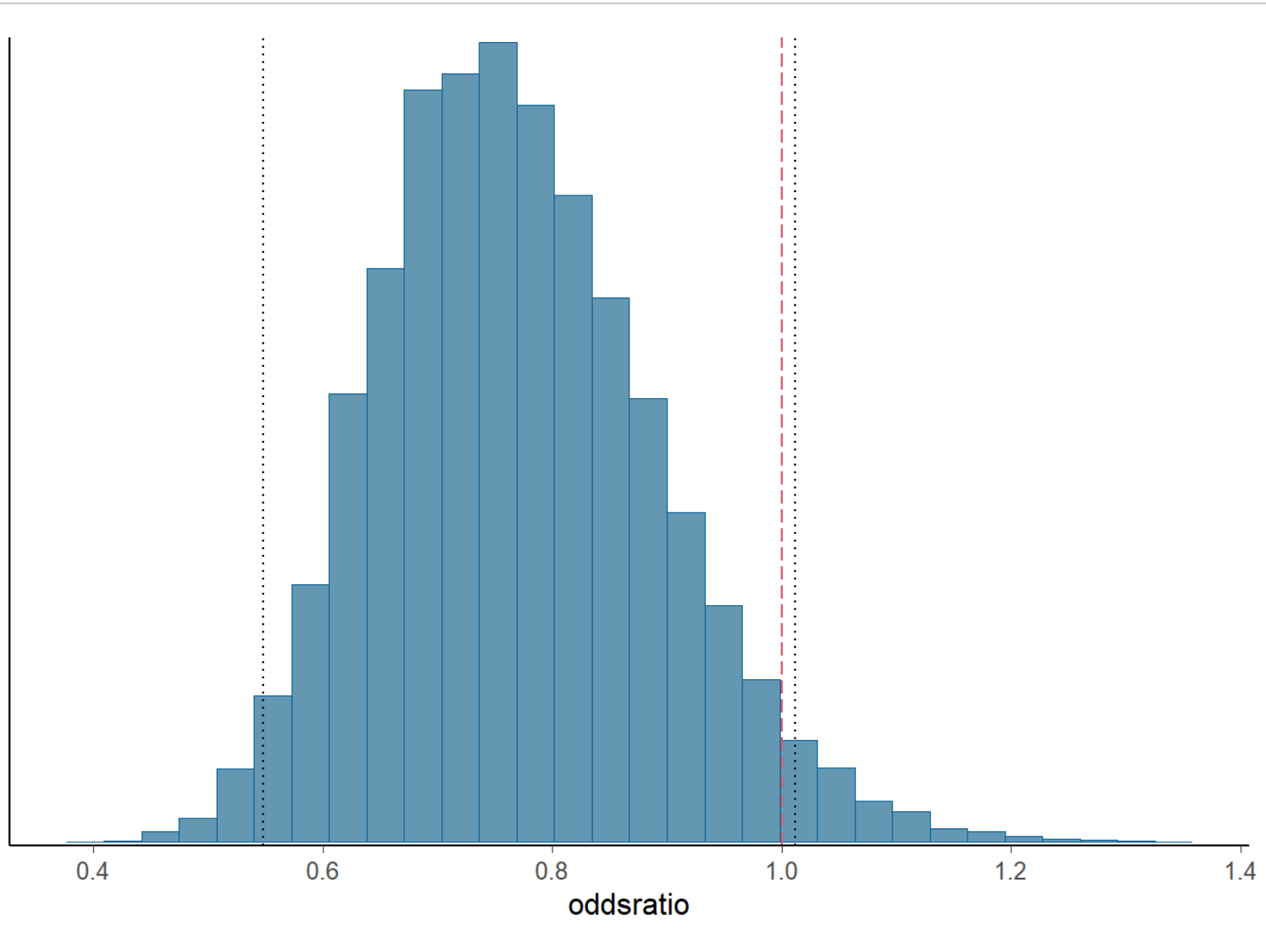
```
## [1] 0.01452714
```

```
CredInts2 = HDInterval::hdi(draws2, 0.95)
CredIntOdds2 = CredInts2[,3]
CredIntOdds2
```

```
##     lower     upper
## 0.5484098 1.0118544
```

```
mcmc_hist(draws2, pars = 'oddsratio') +
  geom_vline(xintercept = CredIntOdds2[1], linetype='dotted') + # plot 95% credible interval
  geom_vline(xintercept = CredIntOdds2[2], linetype='dotted') +
  geom_vline(xintercept = 1, linetype='longdash', col = '2') # plot x = 1 line (odds ratio of thetas equal)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- Mean of Posterior Estimates: 0.7689
- Variance of Posterior Estimates: 0.01453
- Credible Interval of Posterior Estimate: (0.5484, 1.0119)

With the updated priors, we see that the estimate of posterior mean for the odds ratio has increased from 0.7576 to 0.7689. We also see that the 95% credible interval for the odds ratio now includes 1. This is as a result of the more informative prior that we have used in this analysis, which has a greater effect on the shape/distribution of the posterior density compared to the initial uniform prior that was used, making the odds ratio between each of the 2 parameters closer to 1. We can no longer say that the odds ratio will lie below 1 with probability 95%. Thus we cannot conclude that the new drug is more effective than the existing drug with 95% probability.

```
mean(draws2$difference)
```

```
## [1] -0.06778192
```

```
var(draws2$difference)
```

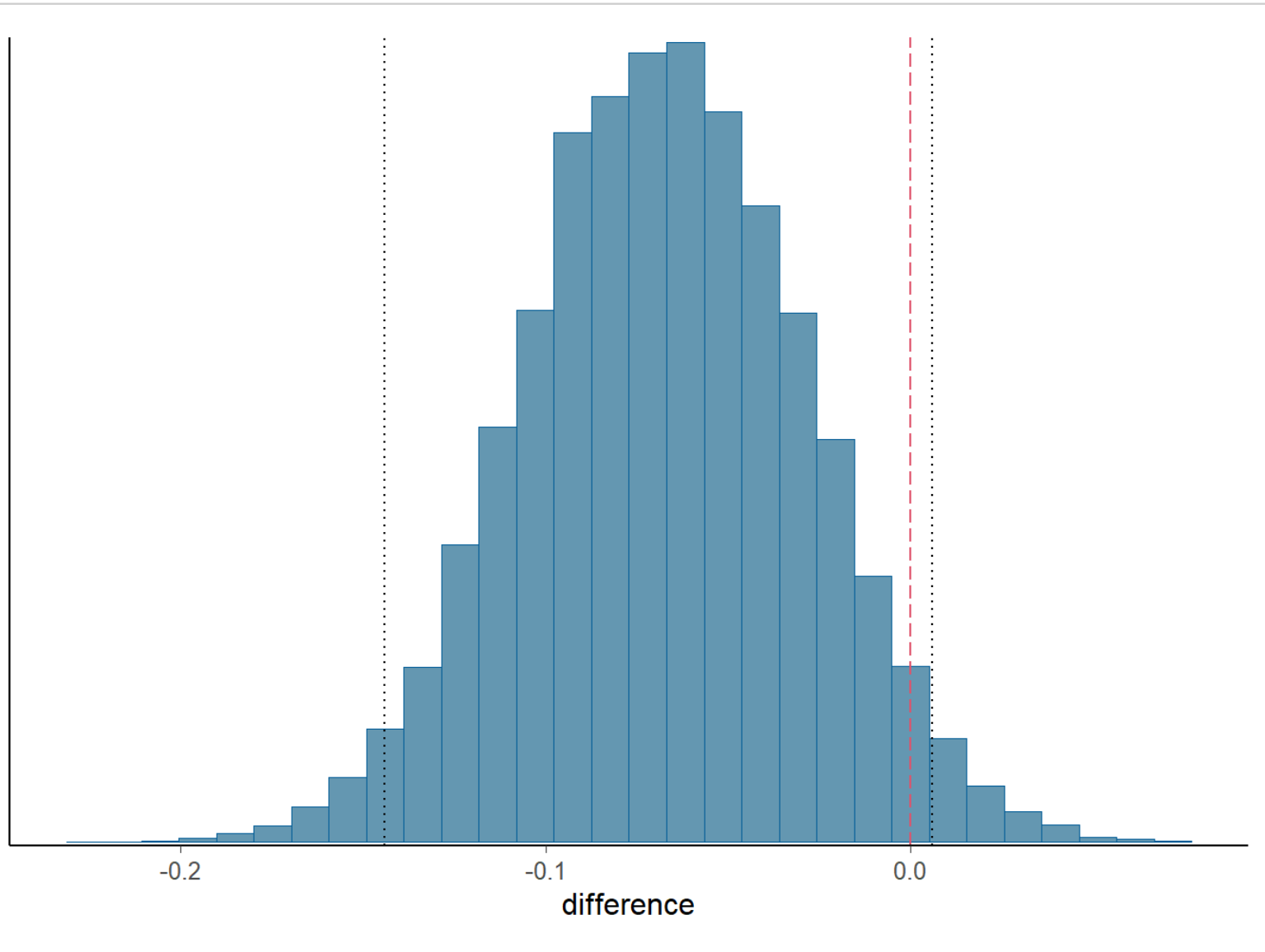
```
## [1] 0.001471734
```

```
# Calculate 95% Credible Interval for Posterior Estimate of Difference
CredIntDiff2 = CredInts2[,4]
CredIntDiff2
```

```
##        lower        upper
## -0.144196939  0.006117001
```

```
mcmc_hist(draws2, pars = 'difference') +
  geom_vline(xintercept = CredIntDiff2[1], linetype='dotted') + # plot the 95% credible interval
  geom_vline(xintercept = CredIntDiff2[2], linetype='dotted') +
  geom_vline(xintercept = 0, linetype='longdash', col = '2') # plot x = 0 line (difference between thetas is 0)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- Mean of Posterior Estimates: -0.0678
- Variance of Posterior Estimates: 0.001472
- Credible Interval of Posterior Estimate: (-0.1442, 0.0061)

Similarly to the odds ratio case, we observe that the updated priors have resulted in an increased posterior mean estimate for the difference between the 2 parameters. The 95% credible interval for posterior distribution for the difference θ2 – θ1 again includes 0 with the updated priors. This means we cannot say the difference in success proportions of the new/existing drugs lies below 0 with 95% probability. Thus we cannot conclude that the new drug is more effective than the existing drug with 95% probability.