EEEB UN3005/GR5005
Homework - Week 05 - Due 05 Mar 2019

Xun Zhao, xz2827

Homework Instructions: Complete this assignment by writing code in the code chunks provided. If required, provide written explanations below the relevant code chunks. Replace "USE YOUR NAME HERE" with your name in the document header. When complete, knit this document within RStudio to generate a pdf. Please review the resulting pdf to ensure that all content relevant for grading (i.e., code, code output, and written explanations) appears in the document. Rename your pdf document according to the following format: hw_week_05_firstname_lastname.pdf. Upload this final homework document to CourseWorks by 5 pm on the due date.

Problem 1 (3 points)

Think back to the research scenario described in last week's homework assignment: you're studying a bacterial pathogen of small mammals, and your pilot sampling efforts have found 9 infected animals out of 20 animals sampled. Much like last week, use grid approximation to construct the posterior for the probability of infection parameter (p). Use 1,000 points in your grid approximation, and assume a flat prior.

Now, let's take our posterior inference a bit further. Generate 10,000 samples from this posterior, assigning them to `samples1`. What is the mean value of `samples1`? What is the 90% HPDI of `samples1`?

```
grid = seq(0, 1, length.out = 1000)
prior = rep(1, 1000)
likelihood = dbinom(9, 20, prob = grid)
unstd.post = likelihood * prior
post = unstd.post / sum(unstd.post)
samples1 = sample(grid, 10000, replace = TRUE, prob = post)
print(mean(samples1))
```

```
## [1] 0.4542986
```

```
print(HPDI(samples1, prob = 0.9))
```

```
##      |0.9      0.9|
## 0.2822823 0.6206206
```

Problem 2 (2 points)

Now imagine that by reading the primary scientific literature, you find that prevalence of this particular bacterial pathogen in similar small mammal populations has never been reported above 0.5. Use grid approximation to construct a posterior distribution for the probability of infection parameter, this time with a prior that assumes that the probability of infection should be < 0.5. Note, a prior that is a constant below $p = 0.5$ and 0 above $p = 0.5$ is a mathematical representation of this assumption.

Generate 10,000 samples from this new posterior distribution (call them `samples2`). What is the mean value of `samples2`? What is the 90% HPDI of `samples2`?

```
prior2 = ifelse(grid < 0.5, 1, 0)
likelihood2 = dbinom(9, 20, prob = grid)
unstd.post2 = likelihood2 * prior2
post2 = unstd.post2 / sum(unstd.post2)
samples2 = sample(grid, 10000, replace = TRUE, prob = post2)
print(mean(samples2))
```

```
## [1] 0.3977448
```

```
print(HPDI(samples2, prob = 0.9))
```

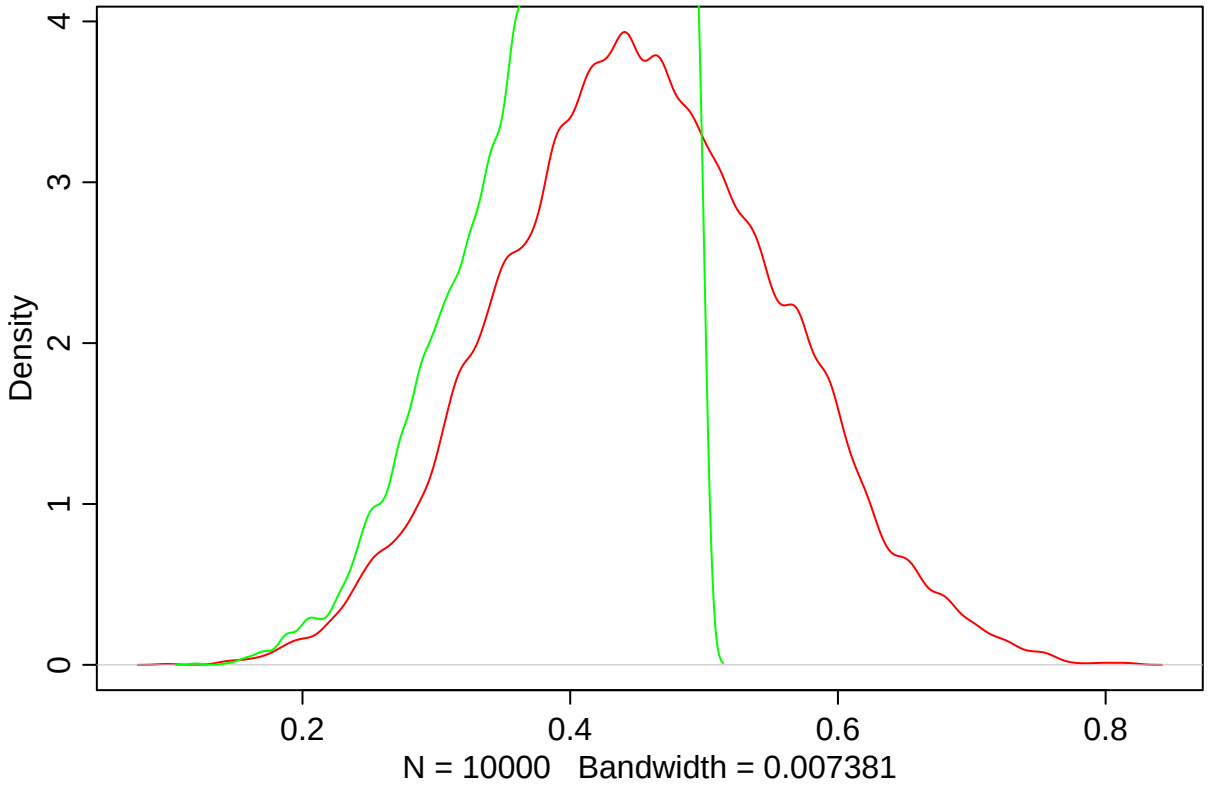
```
##      |0.9      0.9|
## 0.2982983 0.4994995
```

Problem 3 (2 points)

Using the `dens()` function in the `rethinking` package, plot `samples1` and `samples2` together for visual comparison. The `add = TRUE` argument to your second `dens()` call will allow you to plot the samples overlaid. Using different colors for the two different vectors of posterior samples may also help distinguish them. It's just another argument in your `dens()` call (i.e., `col = "red"`).

However you choose to depict them, using this visual comparison as an aid, how does the posterior represented by `samples2` differ from `samples1`? What difference does the change in prior make?

```
dens(samples1, col = 'red')
dens(samples2, add = TRUE, col = 'green')
```



Answer:

The most obvious difference is that `samples2` drops from its maximum to zero at `grid = 0.5`. However, `samples1` shows the symmetrical shape like a Gaussian distribution. At the interval [0, 0.5], `samples2` is larger than `samples1`, but shows the same increasing trend.

The prior makes these differences, because in `likelihood * prior`, new `prior2` makes the second half of the vector all zero, but remains the first half unchanged. That is why `samples2` stops at `0.5`.

Moreover, when standardizing, `sum(unstd.post)` is smaller in `samples2` than `samples1`, which makes `samples2`'s density curve higher than `samples1`'s.

Problem 4 (2 points)

With further literature research, you discover that the true probability of infection is also extremely unlikely to be < 0.2. Modify your grid approximation to include a prior that represents the assumption that $0.2 < p < 0.5$. Generate 10,000 new posterior samples for p using this prior (call these samples `samples3`).

What is the mean value of `samples3`? What is the 90% HPDI of `samples3`?

Does `samples3` differ strongly from `samples2`? Why or why not? As in the previous problem, visual comparison of the posterior distributions may help significantly in answering these questions.

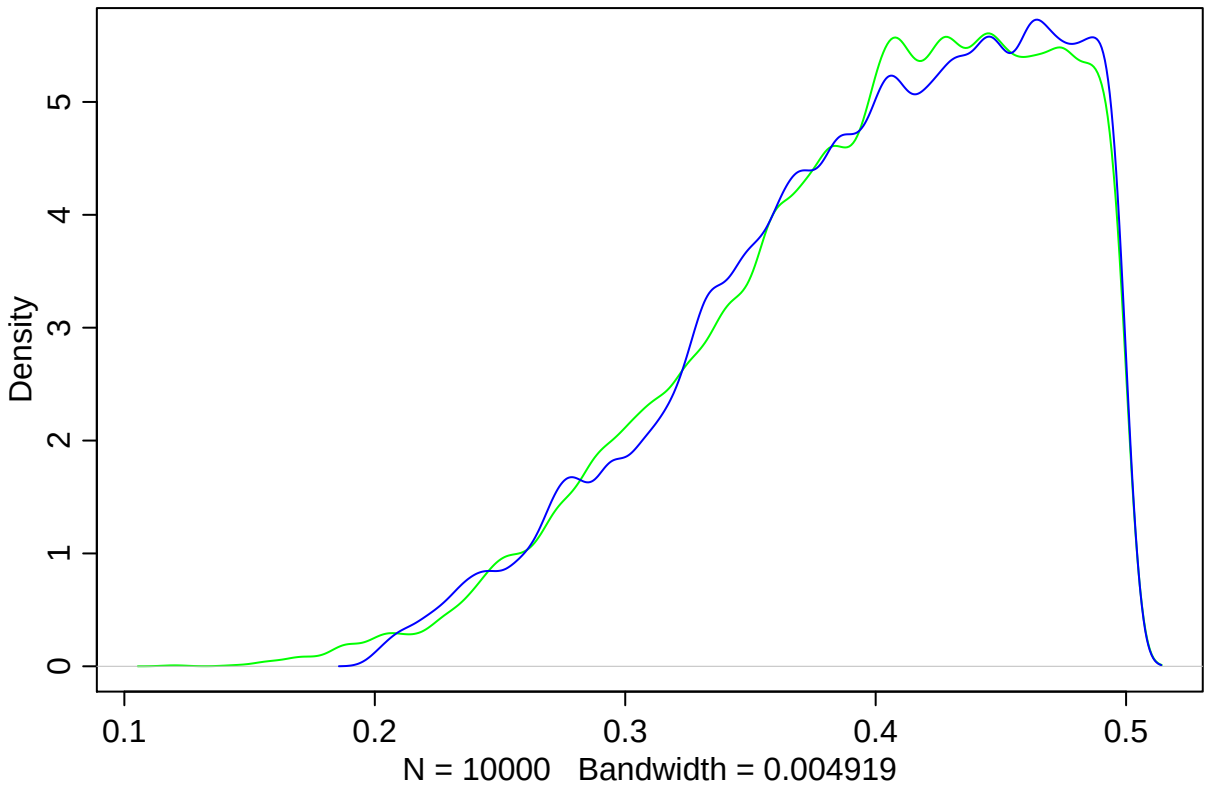
```
prior3 = ifelse(grid < 0.5 & grid > 0.2, 1, 0)
likelihood3 = dbinom(9, 20, prob = grid)
unstd.post3 = likelihood3 * prior3
post3 = unstd.post3 / sum(unstd.post3)
samples3 = sample(grid, 10000, replace = TRUE, prob = post3)
print(mean(samples3))
```

```
## [1] 0.3990964
```

```
print(HPDI(samples3, prob = 0.9))
```

```
##      |0.9      0.9|
## 0.3003003 0.4994995
```

```
dens(samples2, col = 'green')
dens(samples3, add = TRUE, col = 'blue')
```



Answer:

The difference is not significant.

It is because `samples2` originally has low probability to choose from $[0, 0.2]$, as its curve (density) is low when `grid < 0.2`.

Thus, cutting this part from data set makes little impact to the density.

Problem 5 (1 points)

Assume that through further study you establish that the true probability of infection within the population is 0.3. Given this was the true probability of infection value, what was the probability of you initially observing 9 infected individuals out of 20 in your pilot study?

```
probability = dbinom(9, 20, prob = 0.3)
print(probability)
```

```
## [1] 0.06536957
```