Problem Set 3: R Part

Your Name Goes Here

Details

How to Write Up

The written part of this assignment can be either typeset using latex or hand written.

Grading

5% of your grade on this assignment is for turning in something legible. This means it should be organized, and any Rmd files should knit to pdf without issue.

An additional 15% of your grade is for completion. A quick pass will be made to ensure that you've made a reasonable attempt at all problems.

Across both the written part and the R part, in the range of 1 to 3 problems will be graded more carefully for correctness. In grading these problems, an emphasis will be placed on full explanations of your thought process. You don't need to write more than a few sentences for any given problem, but you should write complete sentences! Understanding and explaining the reasons behind what you are doing is at least as important as solving the problems correctly.

Solutions to all problems will be provided.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

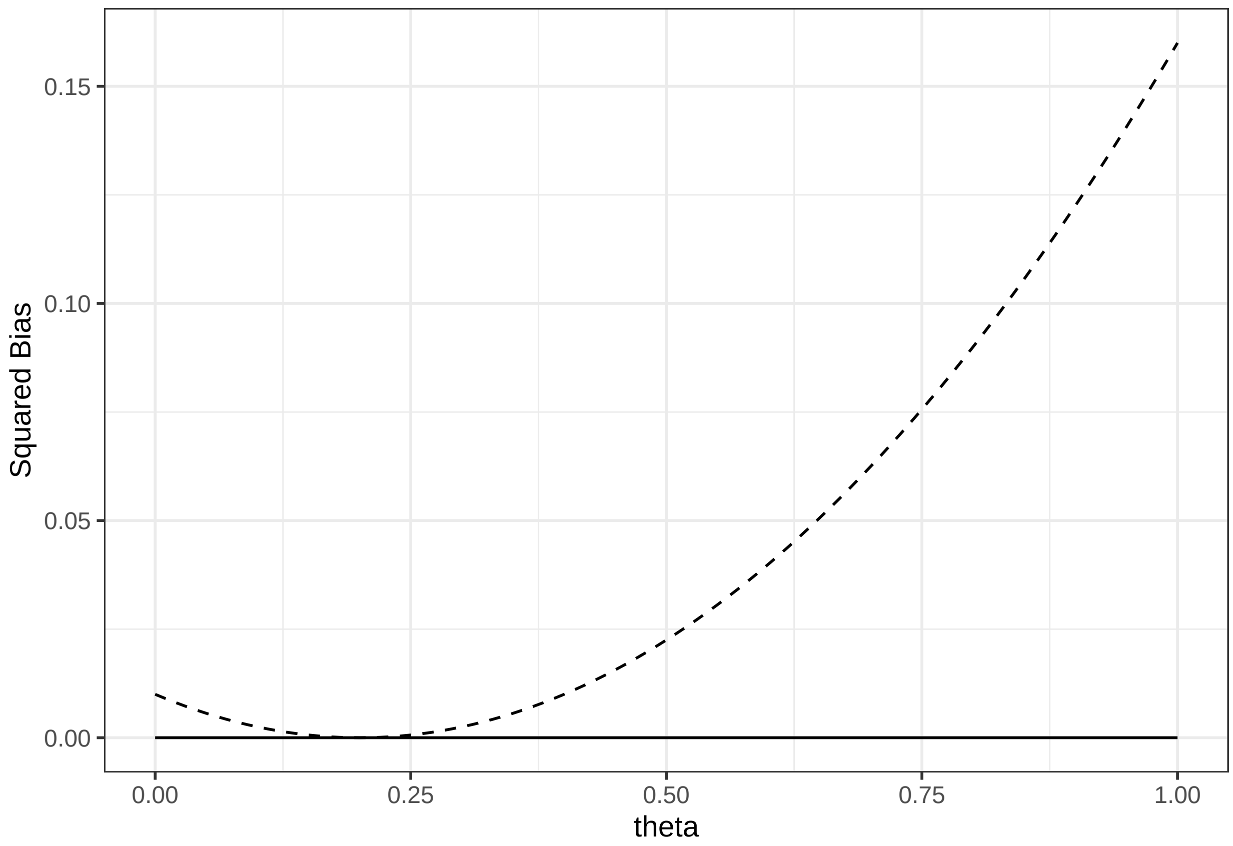
Sources

You may refer to our text, Wikipedia, and other online sources. All sources you refer to must be cited in the space I have provided at the end of this problem set.

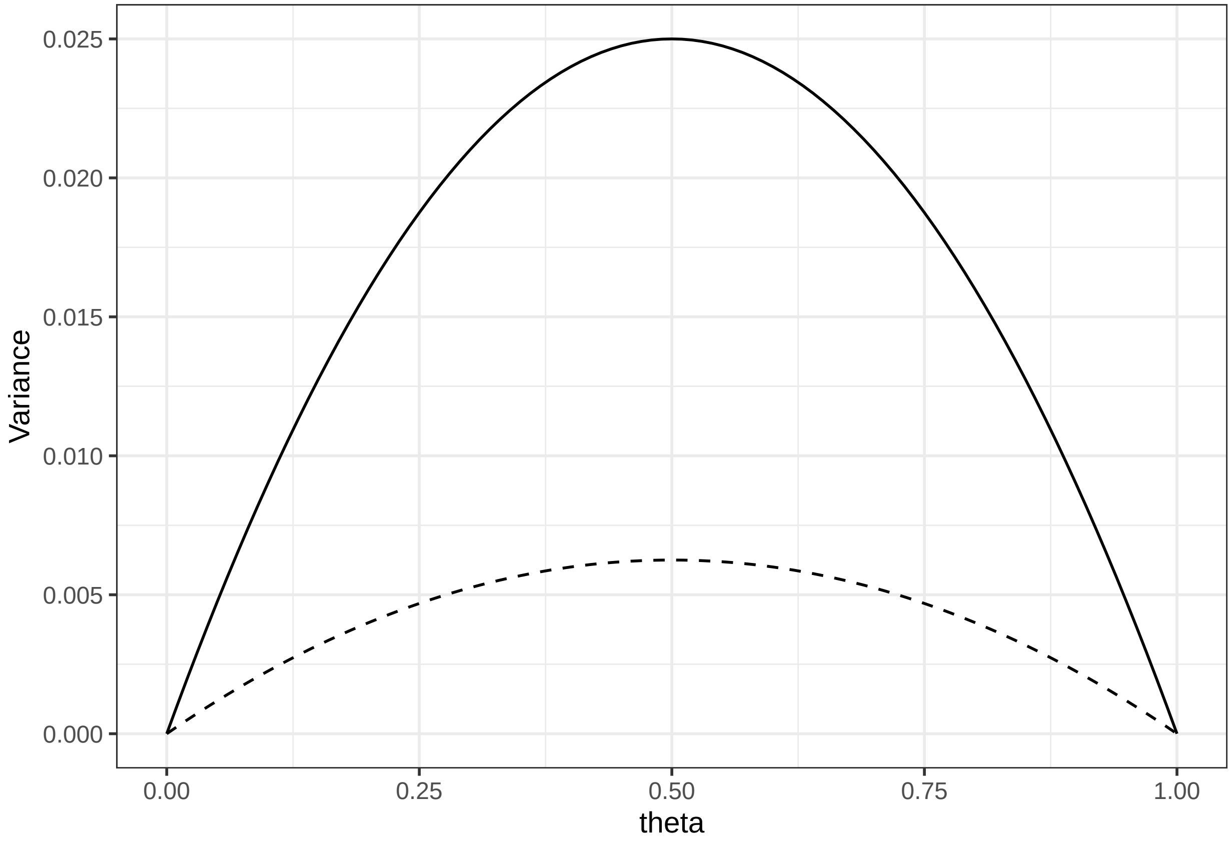
Problem I: M&M's Example: Bias, Variance, Efficiency, and MSE of maximum likelihood estimator and Bayesian posterior mean

(1) Create a series of three plots: the first showing the squared bias of $\hat{\theta}_{MLE}$ and of $\hat{\theta}_{Bayes}$ as a function of θ (the plot's horizontal axis is θ, and it has two curves on it); the second showing the variance of $\hat{\theta}_{MLE}$ and of $\hat{\theta}_{Bayes}$ as a function of θ; and the third showing the MSE of $\hat{\theta}_{MLE}$ and of $\hat{\theta}_{Bayes}$ as a function of θ.

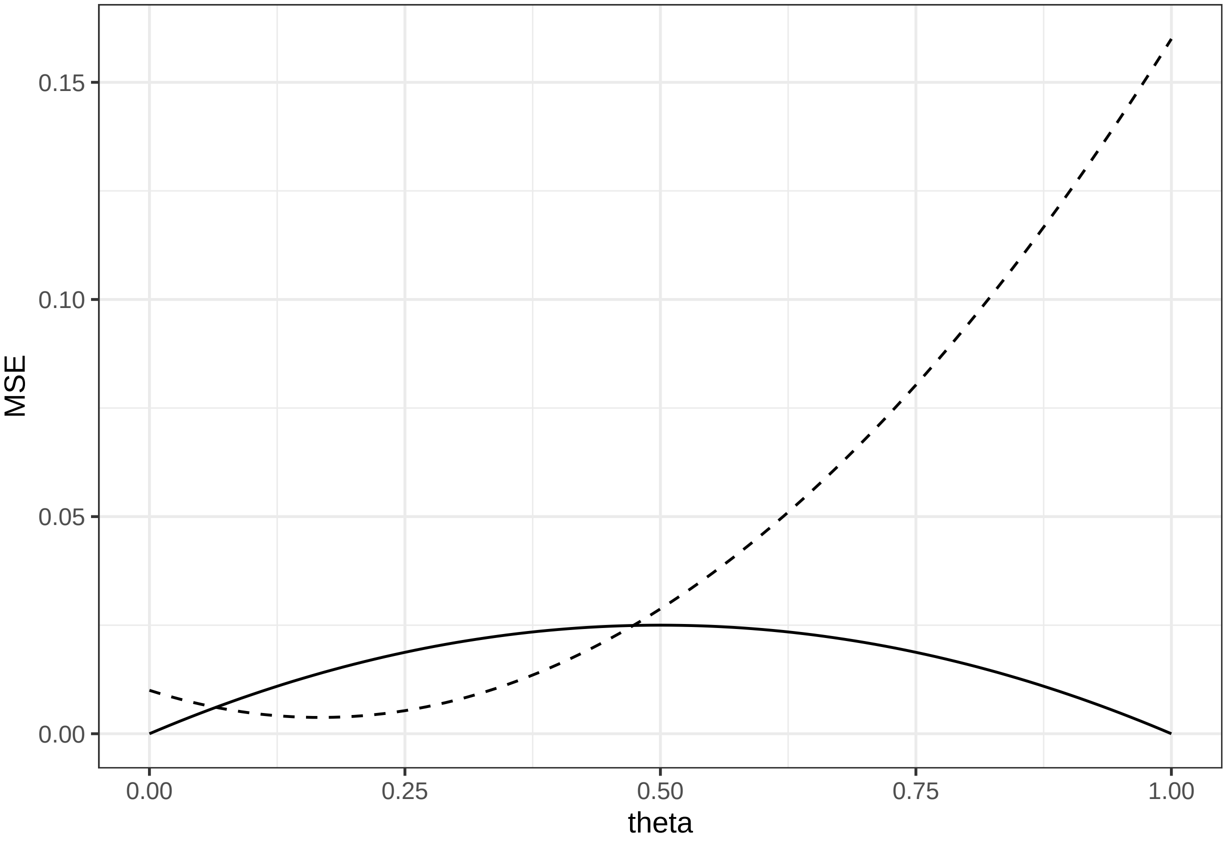
```r
bias_mle <- function(theta) {
  rep(0, length(theta))
}

bias_bayes <- function(theta) {
  (0.5 * 2/10 + 0.5 * theta - theta)^2
}

ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = bias_mle) +
  stat_function(fun = bias_bayes, linetype = 2) +
  ylab("Squared Bias") +
  theme_bw()
```



```r
var_mle <- function(theta) {
  theta * (1 - theta) / 10
}

var_bayes <- function(theta) {
  0.25 * theta * (1 - theta) / 10
}

ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = var_mle) +
  stat_function(fun = var_bayes, linetype = 2) +
  ylab("Variance") +
  theme_bw()
```



```r
mse_mle <- function(theta) {
  theta * (1 - theta) / 10
}

mse_bayes <- function(theta) {
  (0.5 * 2/10 + 0.5 * theta - theta)^2 + 0.25 * theta * (1 - theta) / 10
}

ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = mse_mle) +
  stat_function(fun = mse_bayes, linetype = 2) +
  ylab("MSE") +
  theme_bw()
```



(2) Explain how the plots you created in part 1 relate to each other (which curves add up to which other curves?).

The curves in the plot of MSE are obtained as the sum of the corresponding curves in the plots of squared bias and variance.

(3) For which values of θ (approximately, exact solutions not necessary) is the bias of $\hat{\theta}_{Bayes}$ less than or equal to the bias of $\hat{\theta}_{MLE}$?

$Bias(\hat{\theta}_{Bayes}) \leq Bias(\hat{\theta}_{MLE})$ only if $\theta = 0.2$; this specific value was derived in the written part.

(4) For which values of θ (approximately, exact solutions not necessary) is the variance of $\hat{\theta}_{Bayes}$ less than or equal to the variance of $\hat{\theta}_{MLE}$?

$Var(\hat{\theta}_{Bayes}) \leq Var(\hat{\theta}_{MLE})$ for all values of θ between 0 and 1.

(5) For which values of θ (approximately, exact solutions not necessary) is the MSE of $\hat{\theta}_{Bayes}$ less than or equal to the MSE of $\hat{\theta}_{MLE}$?

$MSE(\hat{\theta}_{Bayes}) \leq MSE(\hat{\theta}_{MLE})$ for values of θ between about 0.06 and 0.45.

(6) Suppose you are actually conducting an analysis where you will estimate θ based on some data, and you will have a sample size of $n = 10$. Which estimator do you prefer to use – $\hat{\theta}_{MLE}$ or $\hat{\theta}_{Bayes}$? There are no correct answers (you can be either frequentist or Bayesian), but you must specifically discuss what the comparisons of bias, variance, and MSE indicate about the relative performance of these estimators and support your preference with specific reference to those comparisons.

There are two possible answers to this question:

1) We prefer the maximum likelihood estimator. The MLE is unbiased, unlike the Bayesian estimator. Although the MLE has larger variance than the Bayesian estimator, if we examine the MSE we find that the MLE has much

more consistent MSE than the Bayesian estimator across the full parameter space. This means that the maximum likelihood estimator will do a reasonable job of estimating θ for any possible value of θ.

2) We prefer the Bayesian estimator. Although the Bayesian estimator is biased, it has a lower variance. When we combine the squared bias and the variance, we find that the Bayesian estimator has lower MSE in a range of plausible values for θ between about 0.065 and 0.47. This means that on average, the Bayesian estimator of θ will have a smaller squared distance from the true value of θ than the maximum likelihood estimator, as long as θ is between about 0.065 and 0.47. Since I am quite sure that the proportion of M&M's that are blue is between 0.065 and 0.47, I prefer the estimator that has best mean squared error in that range of parameter values.

Problem II: Hospital waiting times

The following R code reads the data in:

```
library(readr)
library(ggplot2)
library(dplyr)

er_visits <- read_csv("http://www.evanlray.com/stat343_s2020/homework/ps1/er_wait_times_emergent.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   wait_time = col_double()
## )
```

(1) Suppose you adopt the prior $\Lambda \sim$ Gamma$(1, 0.1)$. Find the parameters of the posterior distribution.

```
nrow(er_visits) + 1
```

```
## [1] 1875
```

```
sum(er_visits$wait_time) + 0.1
```

```
## [1] 69141.1
```

(2) Using the posterior percentiles method, find a 95% posterior credible interval for λ.

You will need to use the `qgamma` function. The parameter `shape` corresponds to your α and `rate` corresponds to your β.

```
qgamma(0.025, shape = nrow(er_visits) + 1, rate = sum(er_visits$wait_time) + 0.1)
```

```
## [1] 0.02590474
```

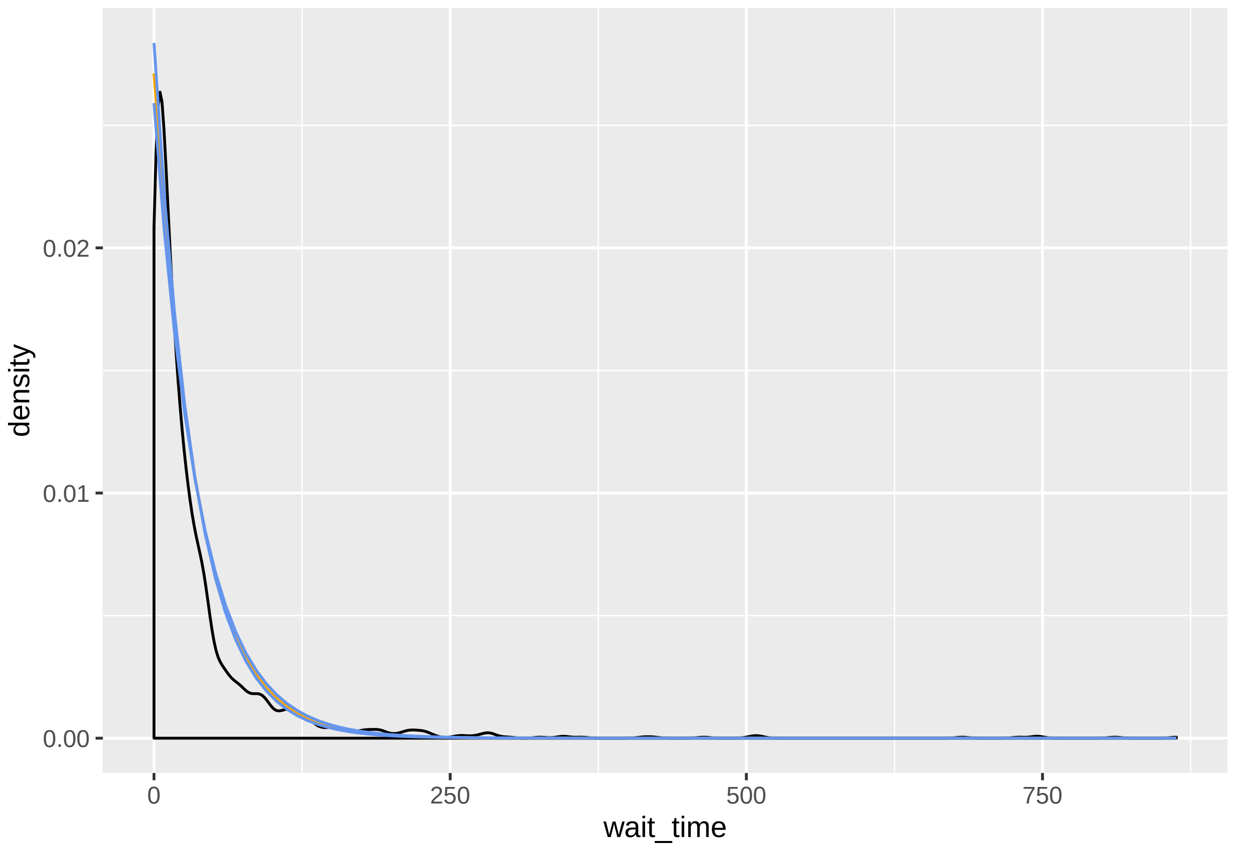
```
qgamma(0.975, shape = nrow(er_visits) + 1, rate = sum(er_visits$wait_time) + 0.1)
```

```
## [1] 0.02835957
```

(3) Add three curves to the plot of the data below showing the pdf of exponential distributions with the posterior mean and the lower and upper endpoints of the 95% posterior credible interval in part (2). Use a color other than black for the densities you add.

The mean of a Gamma(α, β) distribution is $\frac{\alpha}{\beta}$

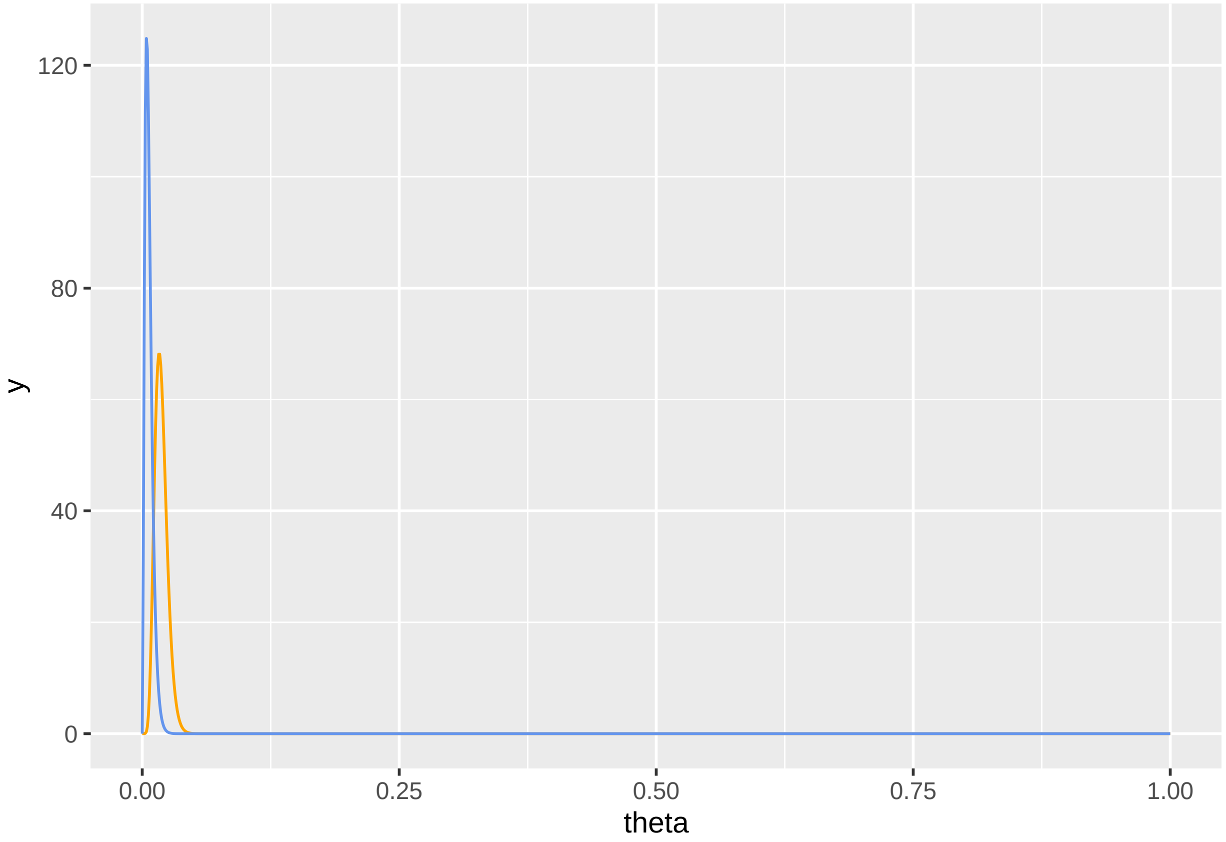
```
ggplot(data = er_visits, mapping = aes(x = wait_time)) +
  geom_density() +
  stat_function(fun = dexp, args = list(rate = (nrow(er_visits) + 1)/(sum(er_visits$wait_time) + 0.1)), color
  stat_function(fun = dexp, args = list(rate = qgamma(0.975, shape = nrow(er_visits) + 1, rate = sum(er_visit
  stat_function(fun = dexp, args = list(rate = qgamma(0.025, shape = nrow(er_visits) + 1, rate = sum(er_visit
```



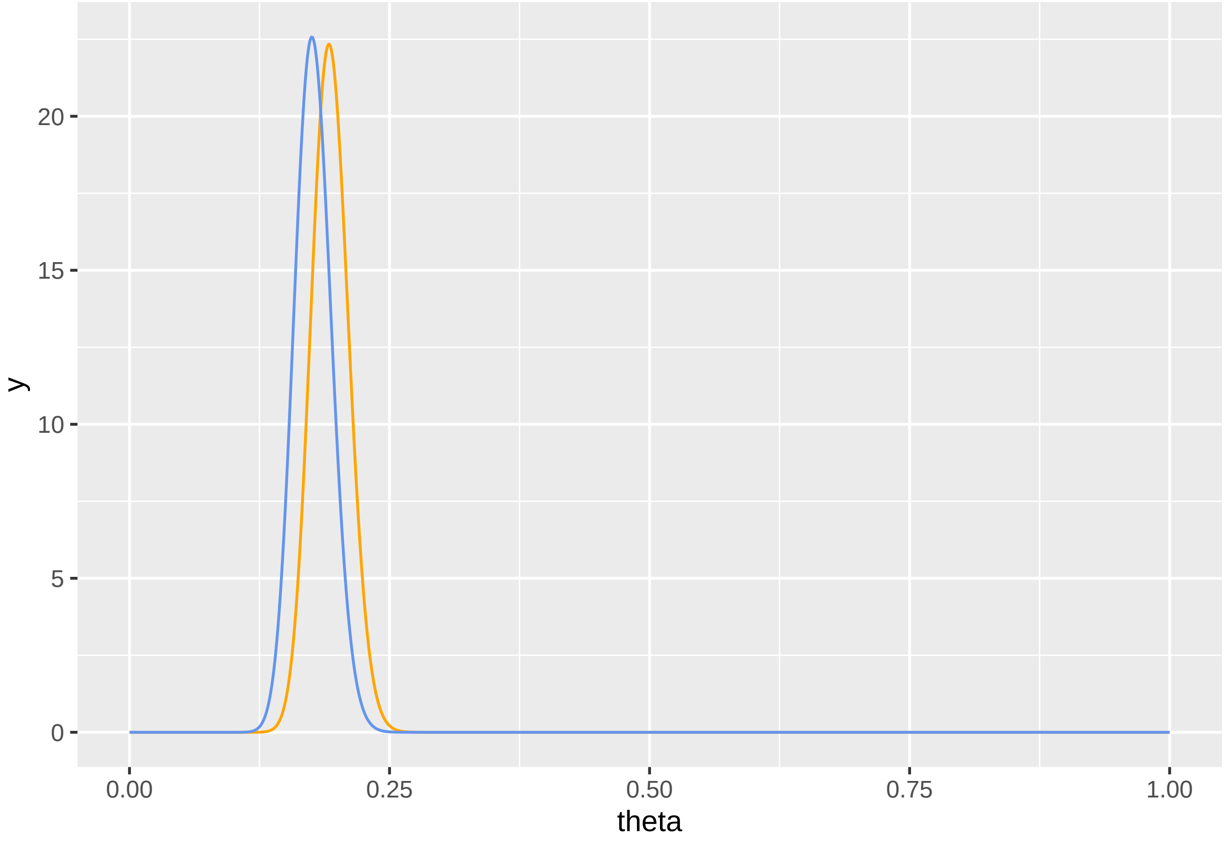
Problem III: Jane Austen

(1) Create a set of 6 plots displaying the marginal posteriors for the model parameters that you found in problem III (3) in the written part. The first of the six plots should show the posteriors for θ_1^{Austen} and $\theta_1^{Imitator}$ i.e., the relative preference for use of "a preceded by such" by each of the two authors. You will then have five more plots for the other parameters.

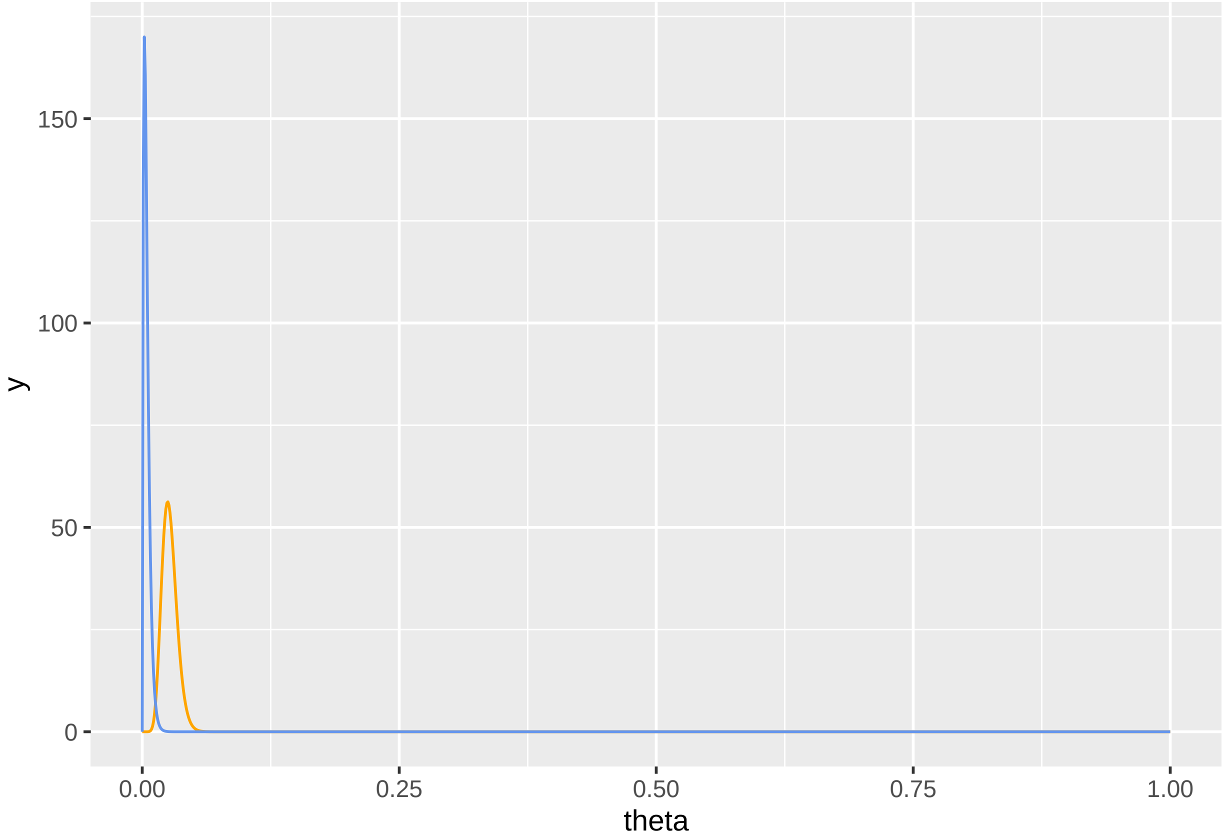
```r
ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta, args = list(shape1 = 9, shape2 = 478), color = "orange", n = 1001) +
  stat_function(fun = dbeta, args = list(shape1 = 3, shape2 = 461), color = "cornflowerblue", n = 1001)
```



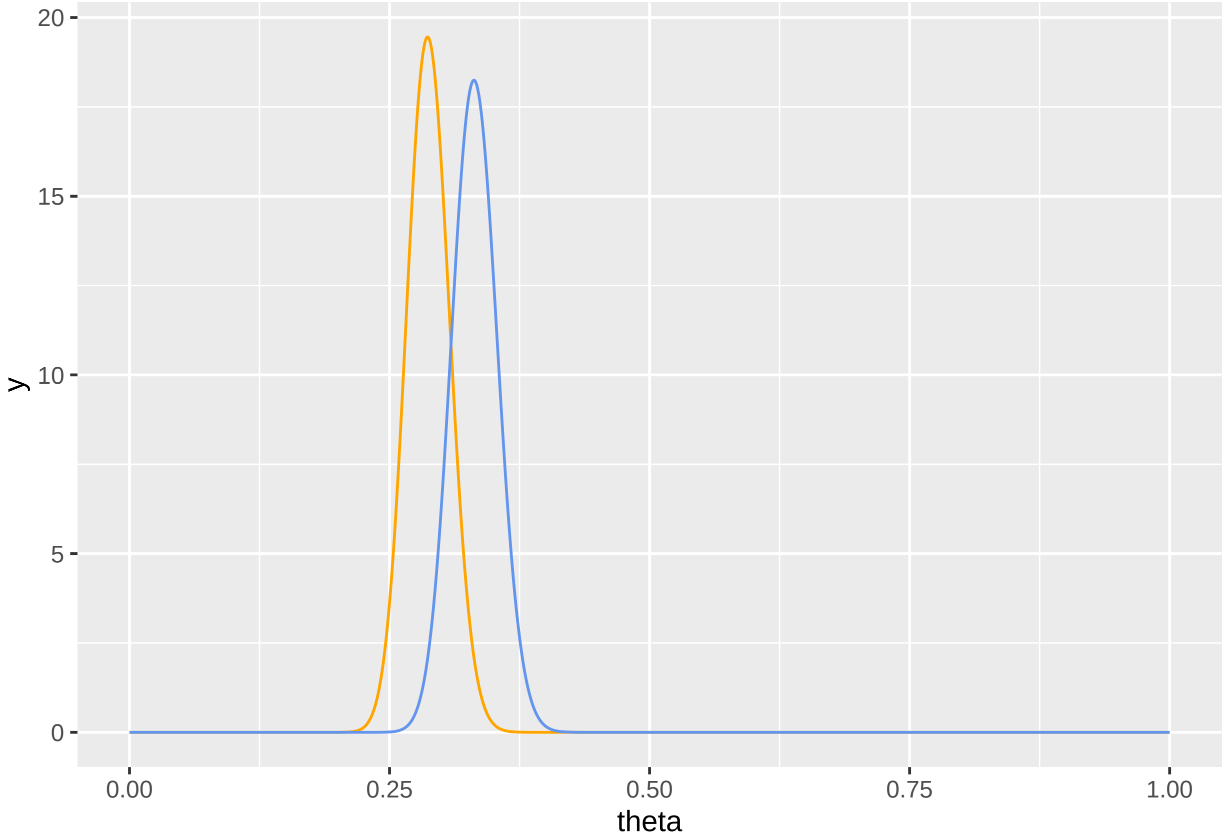
```r
ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta, args = list(shape1 = 94, shape2 = 393), color = "orange", n = 1001) +
  stat_function(fun = dbeta, args = list(shape1 = 82, shape2 = 382), color = "cornflowerblue", n = 1001)
```



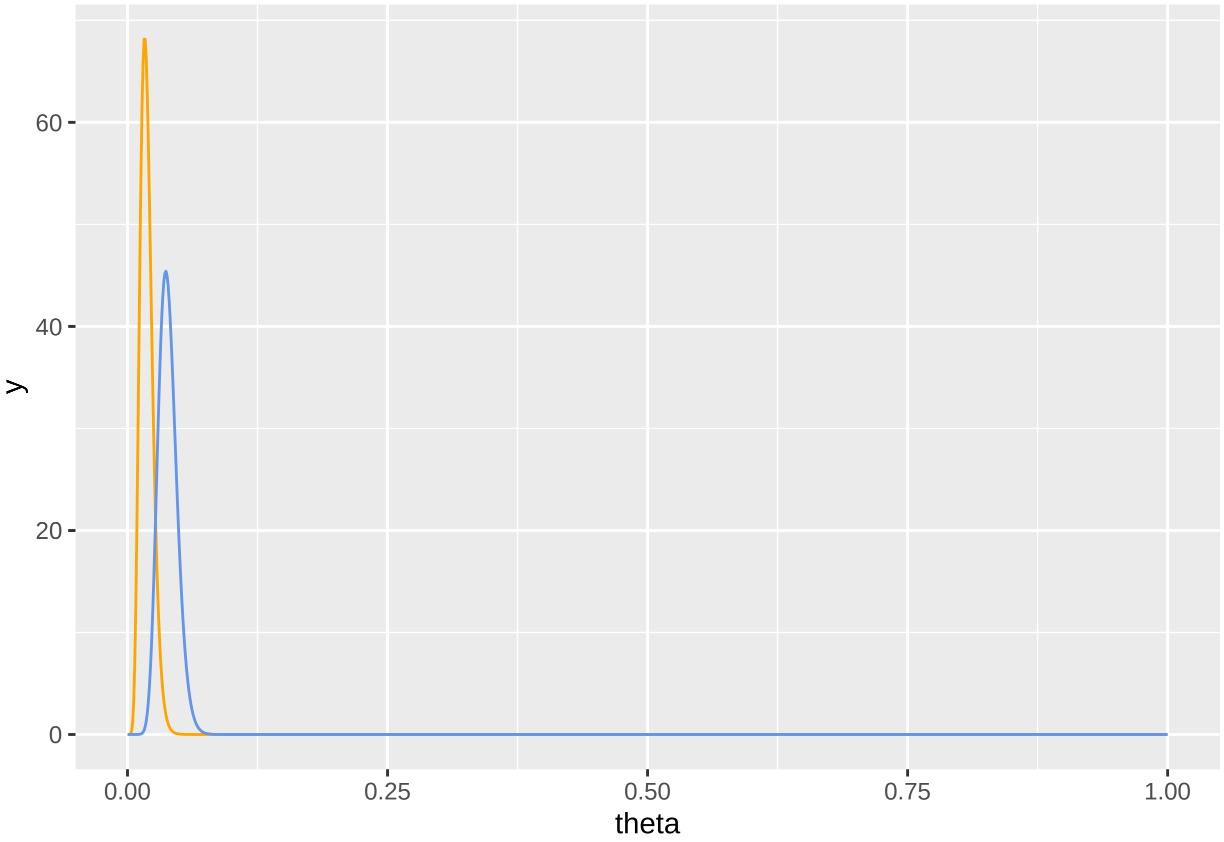
```r
ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta, args = list(shape1 = 13, shape2 = 474), color = "orange", n = 1001) +
  stat_function(fun = dbeta, args = list(shape1 = 2, shape2 = 462), color = "cornflowerblue", n = 1001)
```



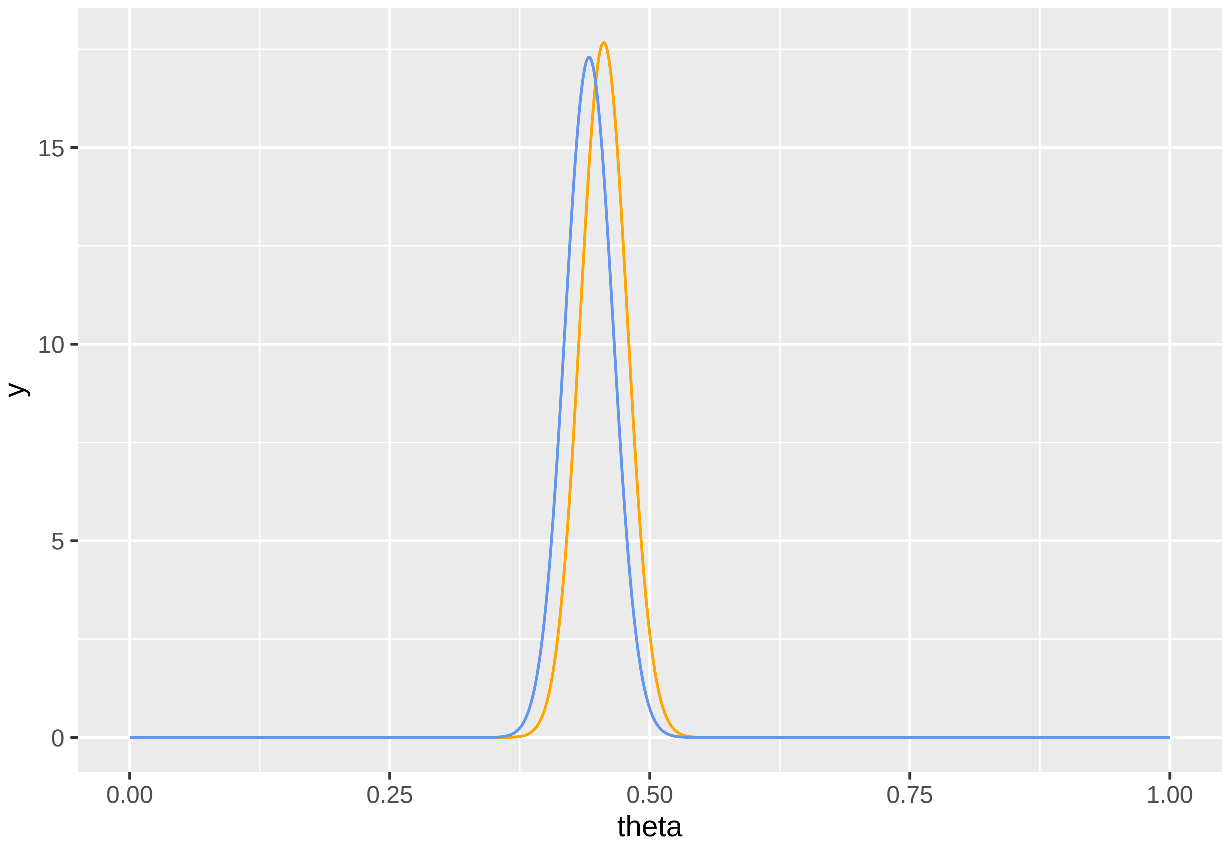
```
ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta, args = list(shape1 = 140, shape2 = 347), color = "orange", n = 1001) +
  stat_function(fun = dbeta, args = list(shape1 = 154, shape2 = 310), color = "cornflowerblue", n = 1001)
```



```
ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta, args = list(shape1 = 9, shape2 = 478), color = "orange", n = 1001) +
  stat_function(fun = dbeta, args = list(shape1 = 18, shape2 = 446), color = "cornflowerblue", n = 1001)
```



```
ggplot(data = data.frame(theta = c(0, 1)), mapping = aes(x = theta)) +
  stat_function(fun = dbeta, args = list(shape1 = 222, shape2 = 265), color = "orange", n = 1001) +
  stat_function(fun = dbeta, args = list(shape1 = 205, shape2 = 259), color = "cornflowerblue", n = 1001)
```



(2) Focus on just the results for "and followed by I" (the third row in the table). Find posterior 95% credible intervals for θ_3^{Austen} and $\theta_3^{Imitator}$. What is the interpretation of your intervals?

```
qbeta(c(0.025, 0.975), shape1 = 13, shape2 = 474)
```

```
## [1] 0.01431790 0.04273334
```

```
qbeta(c(0.025, 0.975), shape1 = 2, shape2 = 462)
```

```
## [1] 0.000523559 0.011974544
```

(3) Based on the displays of the posterior distributions and the credible intervals above, does imitator appear to have done a good job at reproducing characteristics of Jane Austen's writing style? Write 2-3 sentences to explain.

Overall, it appears that the imitator did a remarkably good job at reproducing characteristics of Austen's writing style, but was not perfect. This stands out most in the use of the phrases "a preceeded by such" and "and followed by I", which were used much less often by the imitator than by Austen, as well as the phrase "the preceeded by on", which was used more often by the imitator than by Austen; these observations come from the plots of the posterior distributions as well as the posterior credible intervals. In particular, there is no overlap in the 95% posterior credible intervals for the relative frequency of the use of the phrase "and followed by I" by Austen and the imitator. These differences suggest some systematic differences between Austen's writing and the imitator's writing.