Homework 7

Aakash Kothapally, STA-360

5 PM EDT Friday, November 12

Total points: 10 (reproducibility) + 30 (Q1) = 40 points.

General instructions for homeworks: Please follow the uploading file instructions according to the syllabus. You will give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used. Your code must be completely reproducible and must compile.

Advice: Start early on the homeworks and it is advised that you not wait until the day of. While the professor and the TA's check emails, they will be answered in the order they are received and last minute help will not be given unless we happen to be free.

Commenting code Code should be commented. See the Google style guide for questions regarding commenting or how to write code https://google.github.io/styleguide/Rguide.xml. No late homework's will be accepted.

```
library(MASS)
library(ggplot2)
```

1

. (Multivariate Normal, 30 points, 10 points each) Hoff exercise 7.3 (Australian crab data).

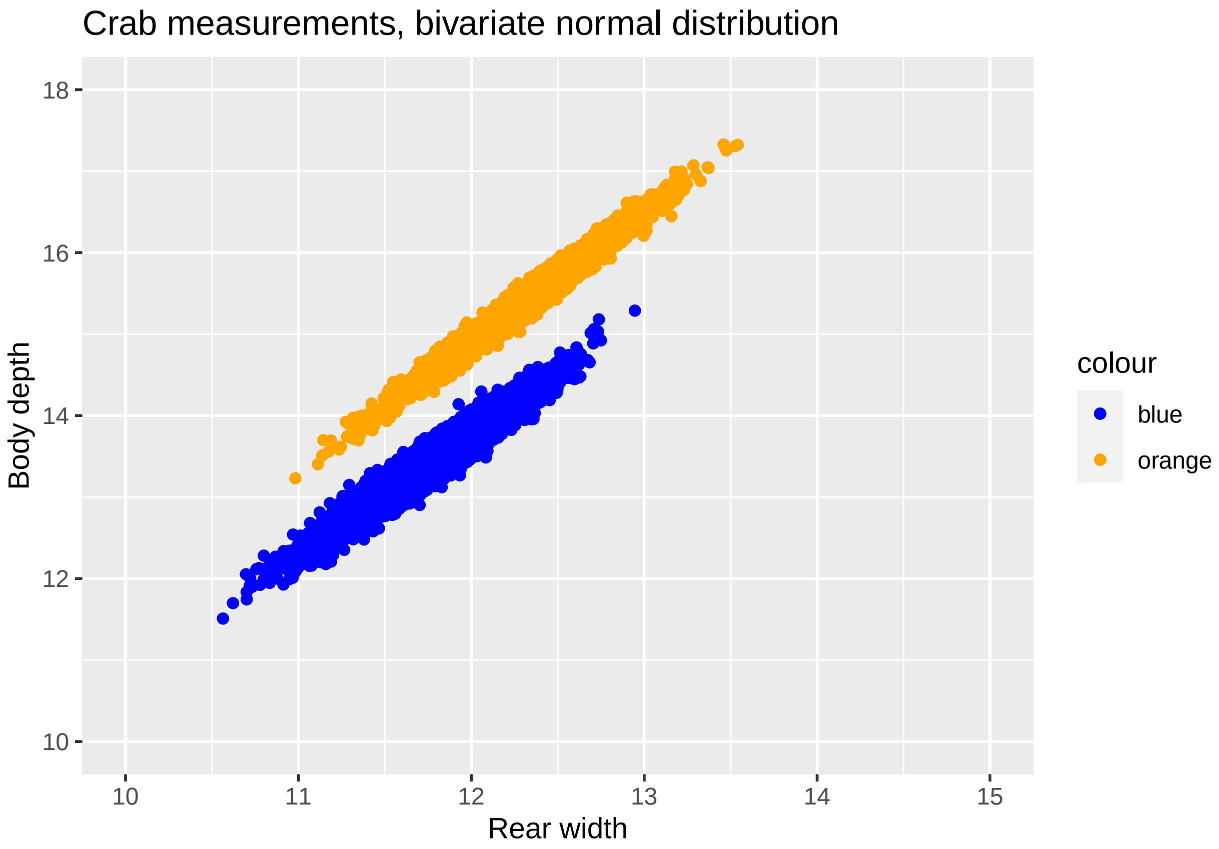
a.

The code below obtains 10,000 posterior samples of theta and sigma.

```
set.seed(42)
bluecrab = as.matrix(read.table('bluecrab.dat'))
orangecrab = as.matrix(read.table('orangecrab.dat'))

multivariatenorm <- function(crab) {
  theta_matrix = matrix(nrow = 10000, ncol = ncol(crab))
  sigma_matrix = array(dim = c(ncol(crab), ncol(crab), 10000))

  lambda_init = cov(crab)
  s_init = cov(crab)
  sigma = cov(crab)

  for (i in 1:10000) {
    lambda_i = solve(solve(lambda_init) + nrow(crab) * solve(sigma))
    mu_i = lambda_i %*% (solve(lambda_init) %*% colMeans(crab) + nrow(crab) * solve(sigma)
                         %*% colMeans(crab))

    resid = t(crab) - c(mvrnorm(n = 1, mu_i, lambda_i))
    sigma = solve(rWishart(1, 4 + nrow(crab), solve(s_init + resid %*% t(resid)))[, , 1])

    theta_matrix[i, ] = mvrnorm(n = 1, mu_i, lambda_i)
    sigma_matrix[, , i] = solve(rWishart(1, 4 + nrow(crab),
                                         solve(s_init + resid %*% t(resid)))[, , 1])
  }

  list(theta = theta_matrix, sigma = sigma_matrix)
}

bluecrab_post <- multivariatenorm(as.matrix(read.table('bluecrab.dat')))
orangecrab_post <- multivariatenorm(as.matrix(read.table('orangecrab.dat')))
```

b.

```
bluecrab_dataframe <- data.frame(bluecrab_post$theta)
colnames(bluecrab_dataframe) = c('theta1', 'theta2')
orangecrab_dataframe <- data.frame(orangecrab_post$theta)
colnames(orangecrab_dataframe) = c('theta1', 'theta2')

ggplot(NULL, aes(x= theta1, y = theta2)) +
  geom_point(data = bluecrab_dataframe, aes(color = 'blue')) +
  geom_point(data = orangecrab_dataframe, aes(color = 'orange')) +
  ggtitle("Crab measurements, bivariate normal distribution") +
  xlim(10, 15) +
  xlab('Rear width') +
  ylab('Body depth') +
  ylim(10, 18) +
  scale_color_identity(guide = "legend")
```



```
mean(orangecrab_dataframe$theta1 > bluecrab_dataframe$theta1)
```

```
## [1] 0.9004
```

```
mean(orangecrab_dataframe$theta2 > bluecrab_dataframe$theta2)
```

```
## [1] 0.9985
```

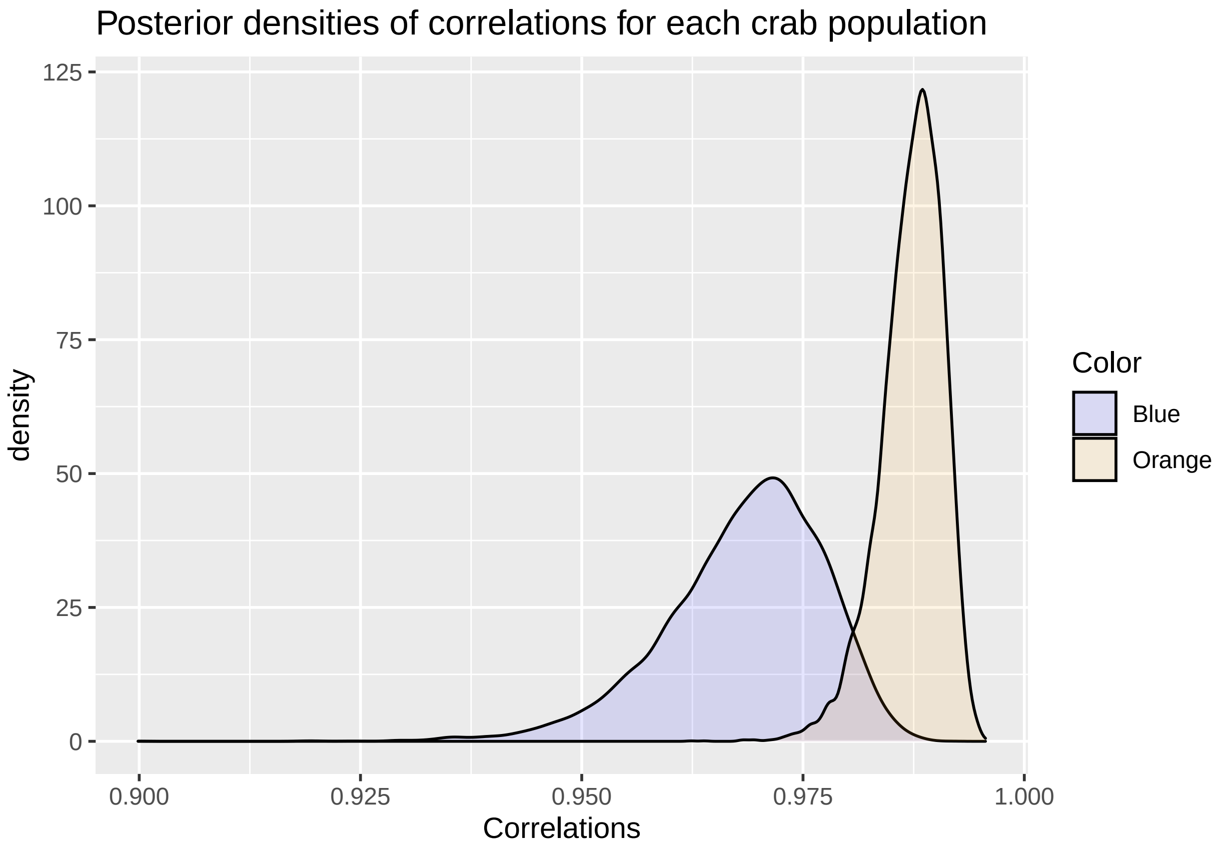
The plot shows on average, there is a positive linear relationship between rear width and body depth for a crab in both the blue crab and orange crab populations.

Also, the plot shows on average, holding body depth constant, the rear widths of orange crabs are 1 mm larger on average, and holding rear width constant, the body depths of orange crabs are 0.5 mm larger on average.

Finally, we can compute the probabilities for both measurements on the orange crabs having larger measurements than blue crabs. These values corroborate our findings.

c.

```
correlation_coeff <- function(covariance_matrix) {covariance_matrix[1, 2] /
    (sqrt(covariance_matrix[1, 1] * covariance_matrix[2, 2]))}

correlation_dataframe = data.frame(
  Color = c(rep('Blue', length(apply(bluecrab_post$sigma, MARGIN = 3, correlation_coeff))),
            rep('Orange', length(apply(orangecrab_post$sigma, MARGIN = 3, correlation_coeff)))),
  Correlations = c(apply(bluecrab_post$sigma, MARGIN = 3, correlation_coeff),
                   apply(orangecrab_post$sigma, MARGIN = 3, correlation_coeff)))

ggplot(correlation_dataframe, aes(x = Correlations, fill = Color)) +
  geom_density(alpha = 0.1) +
  labs(title = "Posterior densities of correlations for each crab population") +
  scale_fill_manual(values = c('Blue', 'Orange'))
```



We can compute an approximation to Pr(pblue > porange | Yblue, Yorange).

```
mean(apply(bluecrab_post$sigma, MARGIN = 3, correlation_coeff)
     < apply(orangecrab_post$sigma, MARGIN = 3, correlation_coeff))
```

```
## [1] 0.9891
```

This tells us the probability is 0.9891 that the orange crab population has a higher correlation between its body depth/rear width measurements than the blue crab population. In other words, the data suggests that the correlation between body depth/rear width is on average higher in orange crabs than blue crabs.