Assignment 4

Kerem Karagöz Immanuel Klein

GitHub: https://github.com/immanuel-klein/bayesian-assignments.git

```
#load packages here
library(dplyr)
library(tidyverse)
library(ggplot2)
library(tinytex)
library(rethinking)
library(rstan)
```

```
# load the data set 'heart.csv' here
heart <- read.csv("heart.csv")
```

Task Set 1

Task 1.1

Run a Bayesian logistic regression model to estimate the risk of men and women to develop a coronary heart disease (TenYearCHD). Provide a summary of the posterior distributions. What is the average probability of men and women to develop the disease?

```
# write data list and model here
heart.chd.gender <- na.omit(heart[, c("male", "TenYearCHD")])

model.gender <- ulam(
  alist(
    TenYearCHD ~ dbinom(1, p),
    logit(p) <- a + bm * male,
    a ~ dnorm(0, 1.5),
    bm ~ dnorm(0, 0.5)
  ), data = heart.chd.gender, chains = 4, cores = 4
)
```

```
#write code here
precis(model.gender, depth = 2)
```

```
         mean         sd       5.5%      94.5%    n_eff    Rhat4
a  -1.9443266 0.06085351 -2.0369612 -1.8433414 854.1949 1.001928
bm  0.4782443 0.08368483  0.3454698  0.6105915 799.1123 1.004692
```

```
samples <- extract.samples(model.gender)
cat("Avg. probability of CHD for women:",
    round(mean(inv_logit(samples$a)), 3))
```

```
Avg. probability of CHD for women: 0.125
```

```
cat("Avg. probability of CHD for men:",
    round(mean(inv_logit(samples$a + samples$bm)), 3))
```

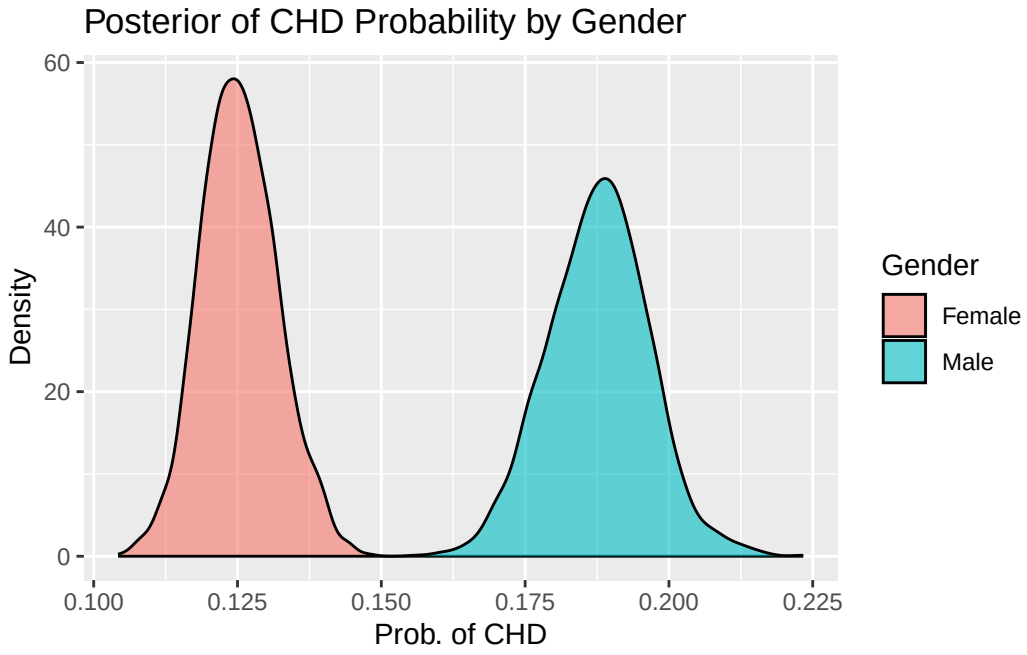
```
Avg. probability of CHD for men: 0.188
```

Task 1.2

For the model of `Task 1.1`, visualize the posterior distribution of gender-differences to assess the credibility of the gender difference.

```
samples.df <- data.frame(
  Female = inv_logit(samples$a),
  Male = inv_logit(samples$a + samples$bm)) %>%
  pivot_longer(cols = c(Female, Male),
               names_to = "Gender", values_to = "Probability")

ggplot(samples.df, aes(x = Probability, fill = Gender)) +
  geom_density(alpha = 0.6) +
  labs(title = "Posterior of CHD Probability by Gender",
       x = "Prob. of CHD",
       y = "Density")
```



Task Set 2

Task 2.1

Run a Bayesian logistic regression model to estimate the risk of men and women with and without diabetes to develop a coronary heart disease (TenYearCHD). Provide a summary of the posterior distributions. Does the effect of diabetes differ between men and women?

```
# write data list and model here
heart.chd.gender.diabetes <- na.omit(
  heart[, c("male", "diabetes", "TenYearCHD")])

model.diabetes <- ulam(
  alist(
    TenYearCHD ~ dbinom(1, p),
    logit(p) <- a + bm * male + bd * diabetes + bmd * male * diabetes,
    a ~ dnorm(0, 1.5),
    bm ~ dnorm(0, 0.5),
    bd ~ dnorm(0, 0.5),
    bmd ~ dnorm(0, 0.5)
  ), data = heart.chd.gender.diabetes, chains = 4, cores = 4
)
```

```
# write code here
# Summarize the posterior distributions
precis(model.diabetes, depth = 2)
```

```
          mean         sd       5.5%      94.5%    n_eff     Rhat4
a   -1.9743969 0.05861428 -2.0659803 -1.8822338 1017.925 0.9984209
bm   0.4643975 0.08285942  0.3310159  0.5974827 1007.401 1.0003998
bd   0.9366253 0.23591513  0.5523493  1.2926296 1231.771 1.0024582
bmd  0.1997384 0.30847011 -0.2842014  0.6983789 1273.817 1.0028843
```

```
samples <- extract.samples(model.diabetes)

cat("Avg. prob. of CHD for women without diabetes:",
    round(mean(inv_logit(samples$a)), 3))
```

```
Avg. prob. of CHD for women without diabetes: 0.122
```

```r
cat("Avg. prob. of CHD for men without diabetes:",
    round(mean(inv_logit(samples$a + samples$bm)), 3))
```

```
Avg. prob. of CHD for men without diabetes: 0.181
```

```r
cat("Avg. probability of CHD for women with diabetes:",
    round(mean(inv_logit(samples$a + samples$bd)), 3))
```

```
Avg. probability of CHD for women with diabetes: 0.264
```

```r
cat("Avg. probability of CHD for men with diabetes:",
    round(mean(inv_logit(samples$a + samples$bm + samples$bd + samples$bmd)),
          3))
```

```
Avg. probability of CHD for men with diabetes: 0.409
```

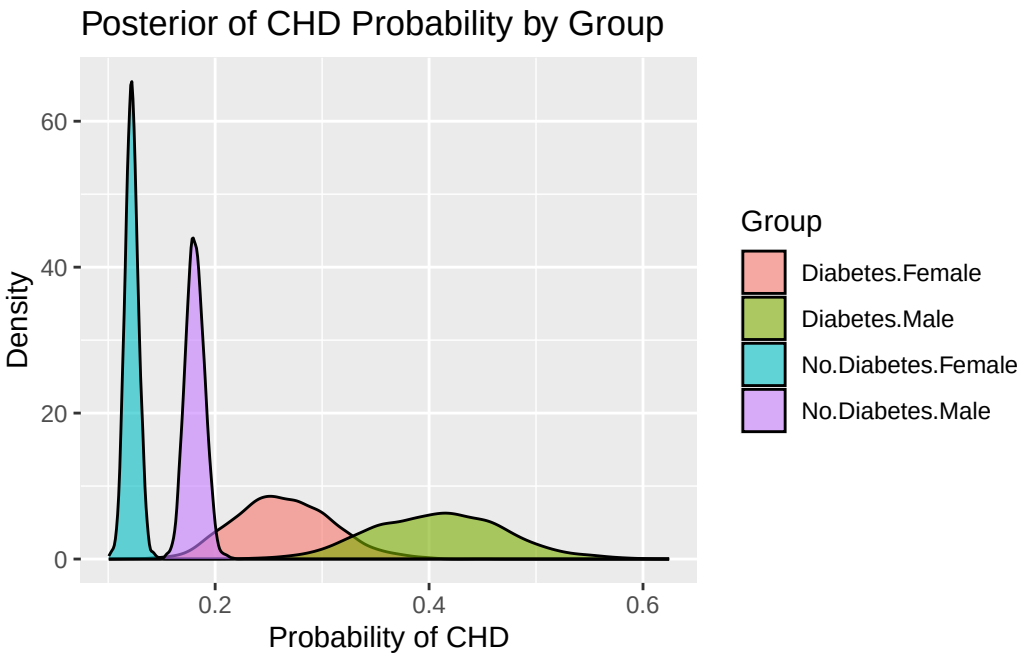
In both cases (diabetes vs. no diabetes) the probability of developing CHD is clearly higher for men than it is for women.

Task 2.2

For the model of `Task 2.1`, visualize the posterior distributions of each group in one plot to better assess the credibility of the group differences.

```r
# write code here
samples.df <- data.frame(
  No.Diabetes.Female = inv_logit(samples$a),
  No.Diabetes.Male = inv_logit(samples$a + samples$bm),
  Diabetes.Female = inv_logit(samples$a + samples$bd),
  Diabetes.Male = inv_logit(samples$a + samples$bm + samples$bd + samples$bmd)
) %>%
  pivot_longer(cols = everything(),
               names_to = "Group",
               values_to = "Probability")

ggplot(samples.df, aes(x = Probability, fill = Group)) +
  geom_density(alpha = 0.6) +
  labs(title = "Posterior of CHD Probability by Group",
       x = "Probability of CHD",
       y = "Density")
```


Posterior of CHD Probability by Group
Density
60
40
20
0
0.2
0.4
0.6
Probability of CHD
Group
Diabetes.Female
Diabetes.Male
No.Diabetes.Female
No.Diabetes.Male

Task Set 3

Task 3.1

Run a Bayesian logistic regression model to estimate the effect of age on the risk of developing a coronary heart disease (TenYearCHD), separately for women and men. Ensure that the regression intercept represents the risk of women and men with average age. Provide a summary of the posterior distributions.

```
# write data list and model here
heart.age <- na.omit(heart[, c("male", "age", "TenYearCHD")])

# Split data by gender and mean center.
# We assume that the data needs to be centered for men and women seperately.
# _ notation instead of . for age_centered,
# because ulam didn't accept age.centered.
heart.male <- heart.age %>% filter(male == 1) %>%
  mutate(age_centered = age - mean(age))
heart.female <- heart.age %>% filter(male == 0)%>%
  mutate(age_centered = age - mean(age))

model.male <- ulam(
  alist(
    TenYearCHD ~ dbinom(1, p),
    logit(p) <- a + bage * age_centered,
    a ~ dnorm(0, 1.5),
    bage ~ dnorm(0, 0.5)
  ), data = heart.male, chains = 4, cores = 4
)

model.female <- ulam(
  alist(
    TenYearCHD ~ dbinom(1, p),
    logit(p) <- a + bage * age_centered,
    a ~ dnorm(0, 1.5),
    bage ~ dnorm(0, 0.5)
  ), data = heart.female, chains = 4, cores = 4
)
```

```
# write code here
precis(model.male, depth = 2)
```

```
            mean         sd        5.5%       94.5%     n_eff    Rhat4
a    -1.56495337 0.06506404 -1.67051654 -1.46459226  985.4453 1.002124
bage  0.06873785 0.00718897  0.05740416  0.08044318 1110.4794 1.002115
```

```r
precis(model.female, depth = 2)
```

```
            mean          sd        5.5%       94.5%    n_eff    Rhat4
a    -2.14350826 0.072635345 -2.26401778 -2.03114911 858.7753 1.002183
bage  0.08472939 0.007931148  0.07195938  0.09737384 835.8826 1.000968
```

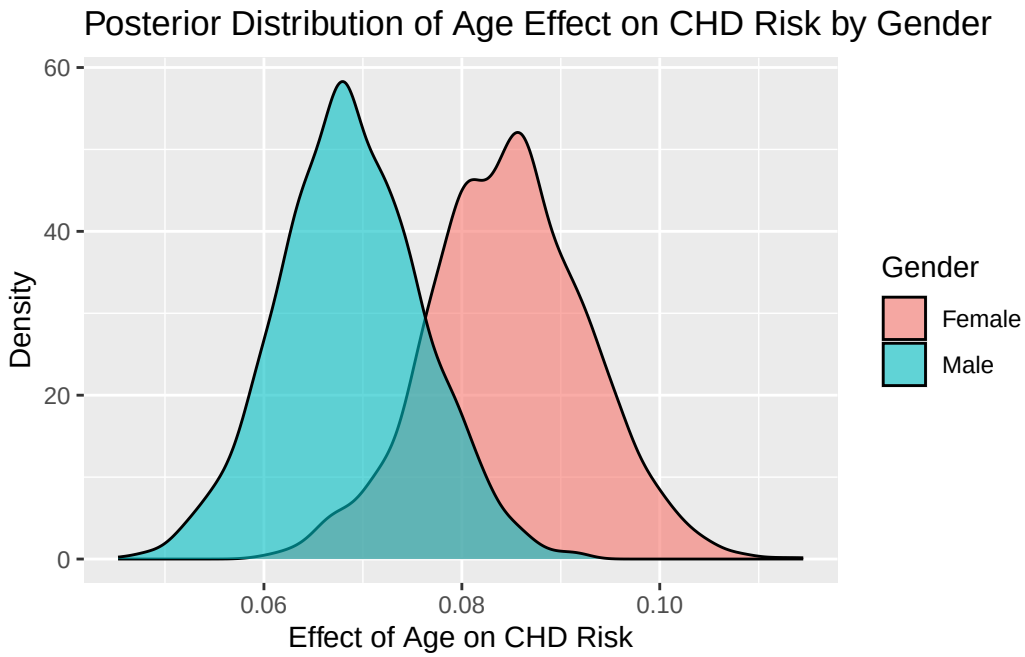
```r
samples.male <- extract.samples(model.male)
samples.female <- extract.samples(model.female)
```

Task 3.2

For the model of `Task 3.1`, visualize the posterior distribution of differences in the age effect between women and men. Does age increase the risk of developing the disease and does this effect differ between women and men?

```r
# write code here
samples.df <- data.frame(
  Male = samples.male$bage,
  Female = samples.female$bage
) %>%
  pivot_longer(cols = c(Male, Female),
               names_to = "Gender",
               values_to = "Age.Effect")

ggplot(samples.df, aes(x = Age.Effect, fill = Gender)) +
  geom_density(alpha = 0.6) +
  labs(title = "Posterior Distribution of Age Effect on CHD Risk by Gender",
       x = "Effect of Age on CHD Risk",
       y = "Density")
```




```
cat("Average age effect for men:", mean(samples.male$bage))
```

```
Average age effect for men: 0.06873785
```

```
cat("Average age effect for women:", mean(samples.female$bage))
```

```
Average age effect for women: 0.08472939
```

Age in fact increases the risk of developing CHD. The influence that age has, is higher for women than it is for men.