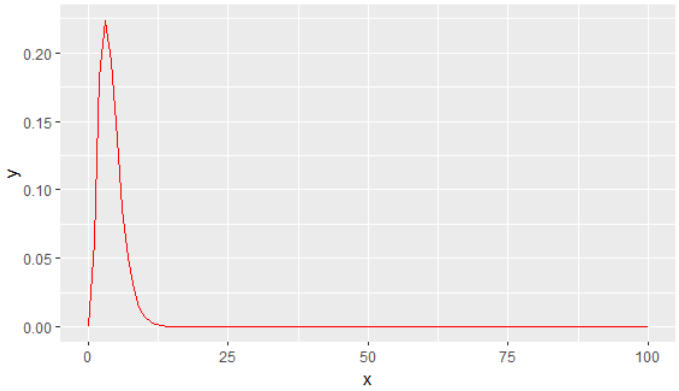
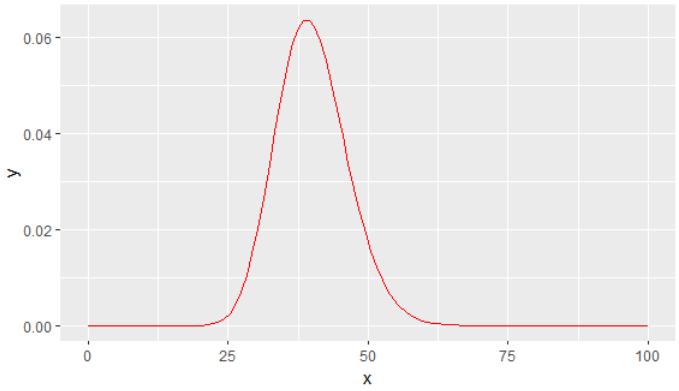
4.8.1

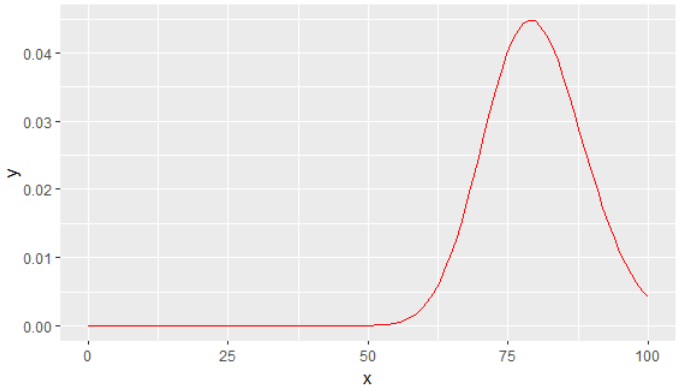
```
x = seq(0,100,length.out = 100)
y = dgamma(x,shape = 4,scale = 1)
data <- data.frame(x,y)
library(ggplot2)
ggplot(data = data, aes(x = x, y = y)) + geom_line(color = "red")
```




```
x = seq(0,100,length.out = 100)
y = dgamma(x,shape = 40,scale = 1)
data <- data.frame(x,y)
library(ggplot2)
ggplot(data = data, aes(x = x, y = y)) + geom_line(color = "red")
```




```
x = seq(0,100,length.out = 100)
y = dgamma(x,shape = 80,scale = 1)
data <- data.frame(x,y)
library(ggplot2)
ggplot(data = data, aes(x = x, y = y)) + geom_line(color = "red")
```



a)

As for these three density functions, the graphs are similar, but the centers of them are different. Among them, the second density function is more symmetric.

b)

The center of the first density function is near the origin; the center of the second density function is approximately in the middle of the graph; the center of the third one is close to the end of the graph.

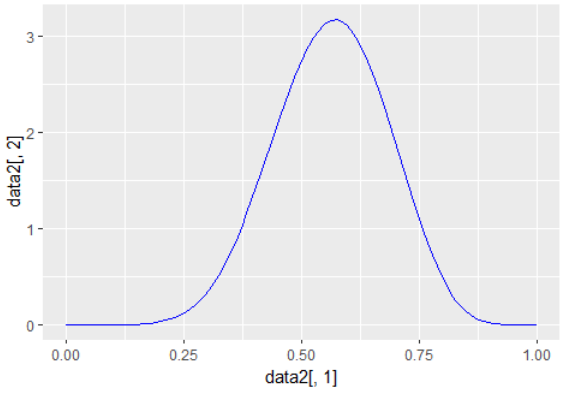
c)

The reason for this difference is that the values of α in these functions are different. As the value of α increases, the center of the graph moves towards the end of the graph.

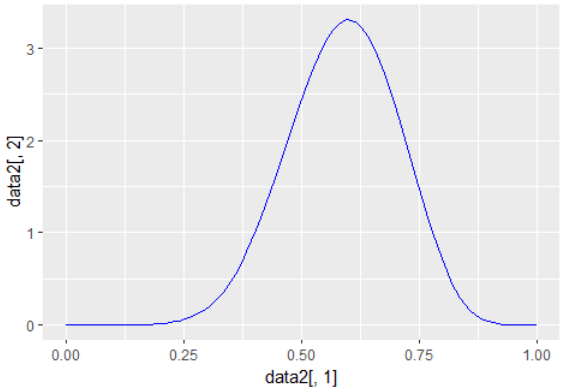
Comment: In this exercise, I learned how to create a sequence of numbers and how to use the function"dgamma".

4.117

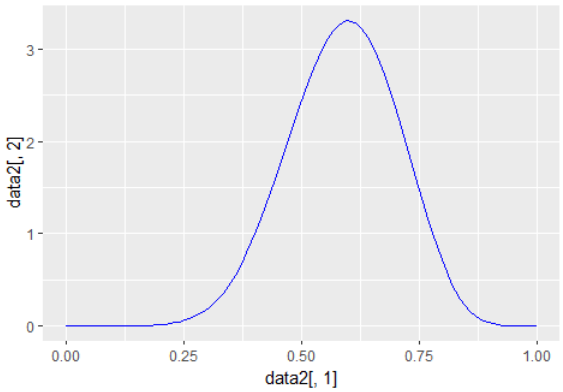
```
x = seq(0, 1, length.out = 100)
y = dbeta(x, shape1 = 9, shape2 = 7)
data2 <- data.frame(x, y)
library(ggplot2)
ggplot(data = data2, aes(x = data2[,1], y = data2[,2])) +geom_line(color = "blue")
```



```
x = seq(0, 1, length.out = 100)
y = dbeta(x, shape1 = 10, shape2 = 7)
data2 <- data.frame(x, y)
library(ggplot2)
ggplot(data = data2, aes(x = data2[,1], y = data2[,2])) +geom_line(color = "blue")
```




```
x = seq(0, 1, length.out = 100)
y = dbeta(x, shape1 = 12, shape2 = 7)
data2 <- data.frame(x, y)
library(ggplot2)
ggplot(data = data2, aes(x = data2[,1], y = data2[,2])) +geom_line(color = "blue")
```



a)

These densities are not symmetric, and all of them are skewed right.

b)

When the value of α gets closer to 12, the center of density function moves towards to the right gradually.

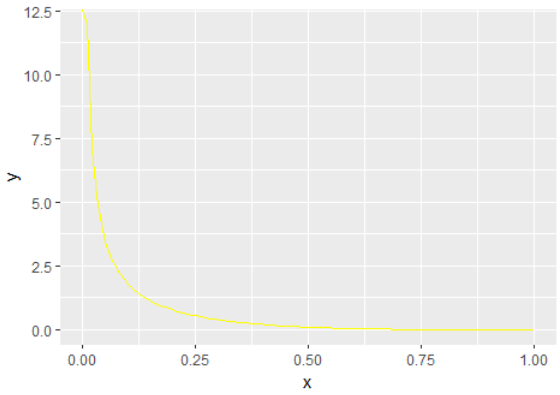
c)

On the condition mentioned in the question, the shapes of beta densities are like the shapes of normal densities. But beta densities' shapes are not symmetric and skewed right.

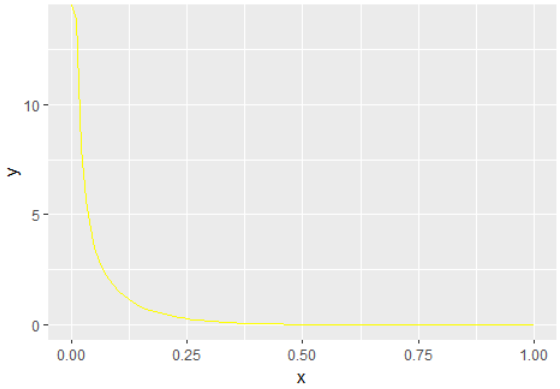
Comment: In this practice, I learned that the ratio of α and β determine the degree to which the beta function skew.

4.118

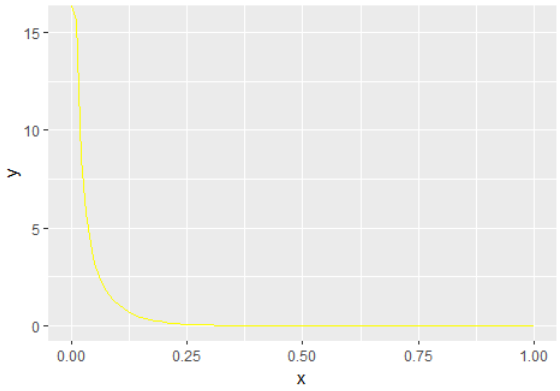
```
x = seq(0, 1, length.out = 100)
y = dbeta(x, shape1 = 0.3, shape2 = 4)
data2 <- data.frame(x, y)
library(ggplot2)
ggplot(data = data2, aes(x = x, y = y)) +geom_line(color = "yellow")
```




```
x = seq(0, 1, length.out = 100)
y = dbeta(x, shape1 = 0.3, shape2 = 7)
data2 <- data.frame(x, y)
library(ggplot2)
ggplot(data = data2, aes(x = x, y = y)) +geom_line(color = "yellow")
```




```
x = seq(0, 1, length.out = 100)
y = dbeta(x, shape1 = 0.3, shape2 = 12)
data2 <- data.frame(x, y)
library(ggplot2)
ggplot(data = data2, aes(x = x, y = y)) +geom_line(color = "yellow")
```



a)
These densities are not symmetric, and all of them are skewed right.

b)
When the value of beta gets closer to 12, the center of the density moves to the left gradually.

c)
The first beta distribution gives the highest probability of observing a value that larger than 0.2

d)
On the condition that beta is smaller than 1, as the value of beta increases the density function moves towards to the right. However, as the value of the α increase, the density function moves towards to the left.

10.19

```
average_x <- 128.6
standard_deviation <- 2.1
z <- (average_x - 130) / standard_deviation
z = -0.67
```

In this test, I used two sides test. After checking the table, I found that when the level is 0.025, the value of Z is 1.96. So, the acceptance region is -1.96 <= Z <= 1.96. -0.67 is in this region, so, I cannot reject the H0 hypothesis.

Comment: In this exercise, I reviewed knowledge that is used for hypothesis testing.

10.21

```
average_y1 <- 1.65
average_y2 <- 1.43
sigma_1 <- 0.26
sigma_2 <- 0.22
Dy <- average_y1 - average_y2
standard_error = sqrt(sigma_1^2 + sigma_2^2)
```

z <- (Dy - 0) / standard_error
z = 0.646
In this test, I used two sides test. After checking the table, I found that when the level is 0.05, the value of Z is 1.64. So, the acceptance region is -1.64 <= Z <= 1.64. Since z = 0.646, this value is in the acceptance region, we cannot reject the H0

Comment: In this exercise, I reviewed how to do hypothesis testing for the difference between two averages. These averages are from two different samples of different populations.

11.31

```
X = c(19.1, 38.2, 57.3, 76.2, 95, 114, 131, 150, 170)
Y = c(0.095, 0.174, 0.256, 0.348, 0.429, 0.500, 0.580, 0.651, 0.722)
average_x <- sum(X) / 9
average_y <- sum(Y) / 9
average_X <- rep(average_x, times = 9)
average_Y <- rep(average_y, times = 9)
x <- X - average_X
y <- Y - average_Y
x <- as.matrix(x)
y <- as.matrix(y)
sum_xy <- t(x) %*% y
sum_x_square <- t(x) %*% x
sum_y_square <- t(y) %*% y
r <- sum_xy / sqrt(sum_x_square * sum_y_square)
t <- (r - 0) / sqrt((1 - r^2) / (9 - 2))
t = 73.04001
```

In this test, I used two sides test. After checking the table, I found that when the level is 0.25, degree of freedom is 7, the value of t is 2,365. That means the acceptance region is -2,365 <= t <= 2,365. Obviously, the 73 is outside this region. Thus, we can reject H0: x and y are not relevant.

11.69

a)

```
Y <- c(18.5, 22.6, 27.2, 31.2, 33.0, 44.9, 49.4, 35.0)
X <- c(-7, -5, -3, -1, 1, 3, 5, 7)
mean_y <- mean(Y)
mean_x <- mean(X)
mean_X <- rep(mean_x, times = 8)
mean_Y <- rep(mean_y, times = 8)
x <- X - mean_X
y <- Y - mean_Y
```

```
vector_xy <- x*y
Sxy <- sum(vector_xy)

x <- as.matrix(x)
Sxx <- t(x) %*% x
beta_1_hat <- Sxy / Sxx
beta_0_hat <- mean_y - beta_1_hat * mean_x
beta_0_hat = 32.7
beta_1_hat = 1.81
```

So, the model Y = 32.7 + 1.81X

b)

```
Y <- c(18.5, 22.6, 27.2, 31.2, 33.0, 44.9, 49.4, 35.0)
X <- c(-7, -5, -3, -1, 1, 3, 5, 7)
vector_1 <- rep(1, times = 8)
vector_3 <- X * X
matrix_x <- cbind(vector_1, X, vector_3)
beta_hat <- solve(t(matrix_x) %*% matrix_x) %*% t(matrix_x) %*% Y
```

beta_hat = [35.56, 1.812, -0.135]

So, the model Y = 35.56 + 1.812X -0.135X^2

2.1

1. Which type of missing mechanism do you prefer to get a good imputation?

I would like to impute using" Imputation based on logical rules". This is because using this method can help me to recover the missing values as exactly as possible.

2. Say something about simple random imputation and regression imputation of a single variable.

Simple random imputation means if there are some missing data in one variable from some units, we can just randomly select some values of this variable from other units and replace NA with the values we have selected.

Regression imputation means we use the units without missing data to create a linear model, which can be used to predict the outcomes with predictors. In this way, these missing values of a certain variable can be predicted by using other variables of these units as inputs.

3. Explain shortly what Multiple Imputation is.

In fact, we seldom have a dataset with only one variable having missing data. In this case, we have to find some methods to predict these missing data in different multiple variables. The process to deal with this situation is called Multiple Imputation.