Stats 102B HW1

2023-04-13

Problem 1

```
library(MASS)
```

a

```
samp_size = 20
r = 0
set.seed(100)
mvrnorm(n = samp_size, mu = c(0,0), Sigma = matrix(data = c(1,r,r,1), nrow= 2, byrow = TRUE))
```

```
##              [,1]        [,2]
##  [1,]  0.43808998 -0.50219235
##  [2,] -0.76406062  0.13153117
##  [3,] -0.26196129 -0.07891709
##  [4,] -0.77340460  0.88678481
##  [5,]  0.81437912  0.11697127
##  [6,]  0.43845057  0.31863009
##  [7,]  0.72022155 -0.58179068
##  [8,] -0.23094453  0.71453271
##  [9,]  1.15772946 -0.82525943
## [10,] -0.24707599 -0.35986213
## [11,]  0.09111356  0.08988614
## [12,] -1.75737562  0.09627446
## [13,]  0.13792961 -0.20163395
## [14,]  0.11119350  0.73984050
## [15,]  0.69001432  0.12337950
## [16,]  0.22179423 -0.02931671
## [17,] -0.18290768 -0.38885425
## [18,] -0.41732329  0.51085626
## [19,] -1.06540233 -0.91381419
## [20,] -0.97020202  2.31029682
```

b

Generate simulated data from a bivariate multivariate normal distribution with mean vector mu = [00] and correlation matrix R for the following cases: ### 1. Sample size n in {20, 50, 100, 200} and correlation coefficient r = 0

```
set.seed(100)
sample_list_part1 <- list(b1_sample20 = matrix(nrow = 20, ncol = 2),
                    b1_sample50 = matrix(nrow = 50, ncol = 2),
                    b1_sample100 = matrix(nrow = 100, ncol = 2),
                    b1_sample200 = matrix(nrow = 200, ncol = 2))
```

```
samp_size = c(20,50,100,200)
r = 0
for(i in seq_along(samp_size)){
  sample_list_part1[[i]]<- mvrnorm(n = samp_size[i], mu = c(0,0), Sigma = matrix(data = c(1,r,r,1), nrou
}
```

2. Sample size n in {20, 50, 100, 200} and correlation coefficient r = 0.5

```
set.seed(100)
sample_list_part2 <- list(b1_sample20 = matrix(nrow = 20, ncol = 2),
                    b1_sample50 = matrix(nrow = 50, ncol = 2),
                    b1_sample100 = matrix(nrow = 100, ncol = 2),
                    b1_sample200 = matrix(nrow = 200, ncol = 2))


samp_size = c(20,50,100,200)
r = 0.5
for(i in seq_along(samp_size)){
  sample_list_part2[[i]]<- mvrnorm(n = samp_size[i], mu = c(0,0), Sigma = matrix(data = c(1,r,r,1), nrou
}
```

3. Sample size n in {20, 50, 100, 200} and correlation coefficient r = 0.85

```
set.seed(100)
sample_list_part3 <- list(b1_sample20 = matrix(nrow = 20, ncol = 2),
                    b1_sample50 = matrix(nrow = 50, ncol = 2),
                    b1_sample100 = matrix(nrow = 100, ncol = 2),
                    b1_sample200 = matrix(nrow = 200, ncol = 2))


samp_size = c(20,50,100,200)
r = 0.85
for(i in seq_along(samp_size)){
  sample_list_part3[[i]]<- mvrnorm(n = samp_size[i], mu = c(0,0), Sigma = matrix(data = c(1,r,r,1), nrou
}
```

c

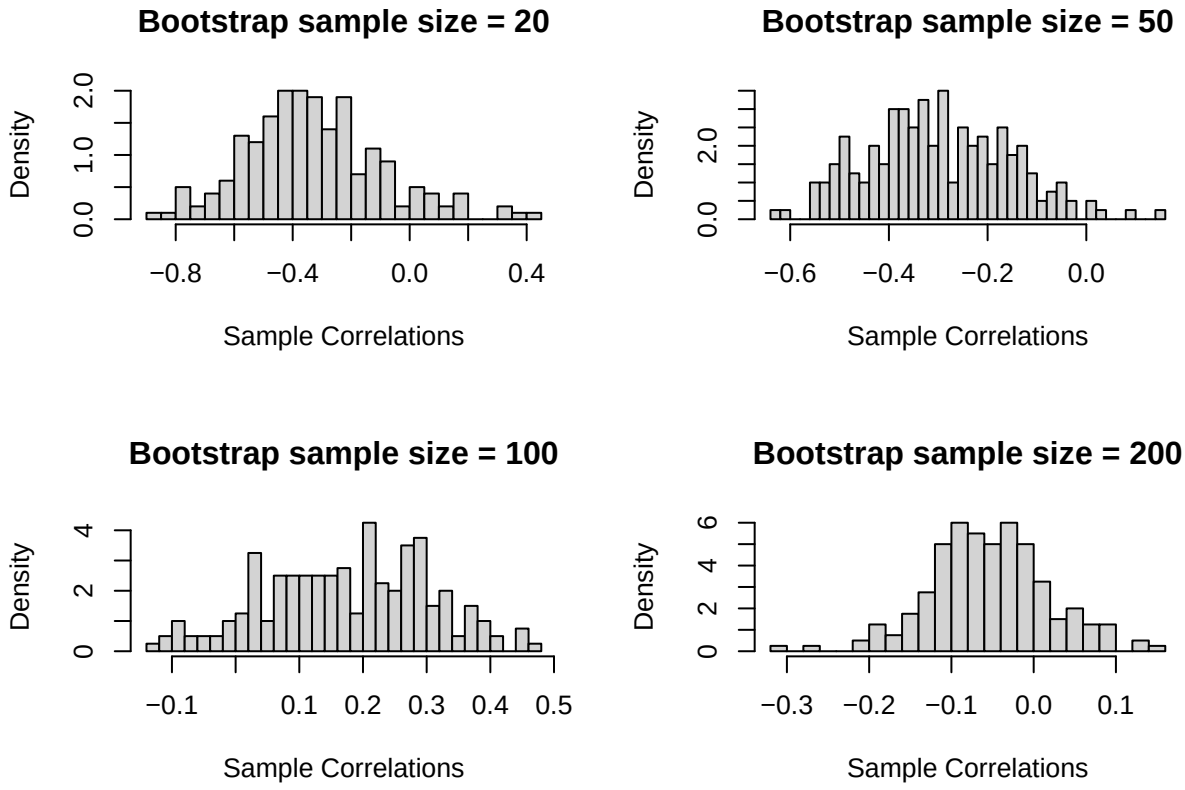
Obtain the bootstrap sampling distribution of the sample correlation coefficient r^ for the three cases in part (b), for the following number of bootstrap replicates B in {200, 1000, 5000, 10000}. Comment on the results; in particular how the bootstrap sampling distribution behaves as a function of the sample size n, the number of bootstrap replicates B and the value of the correlation coefficient r.

Case 1

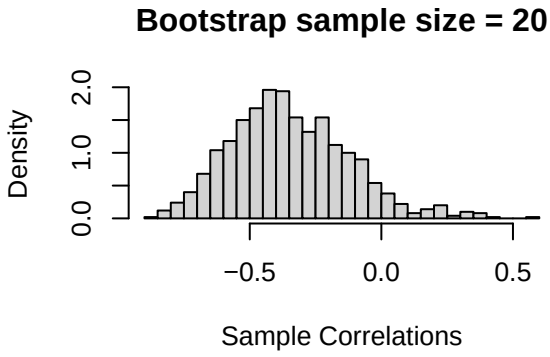
```
bootsampling <- function(x, boot.replicates = B){
  x = as.matrix(x)
  nx = nrow(x)
  return(replicate(boot.replicates, x[sample.int(nx, replace = TRUE),]))
}
```

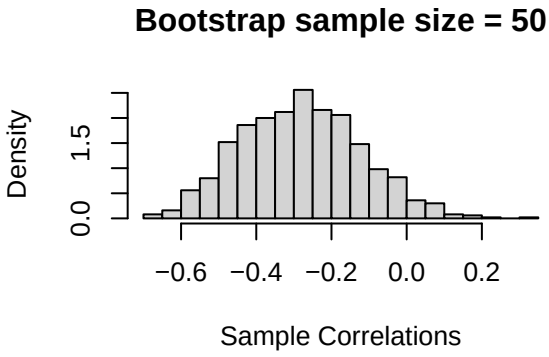
```r
set.seed(100)
bootstrap_replicates = c(200, 1000, 5000, 10000)

#B = bootstrap_replicates
for(B in seq_along(bootstrap_replicates)){
  print(paste("Bootstrapping with B= ", bootstrap_replicates[B]))
  par(mfrow = c(2,2))
  for(i in seq_along(samp_size)){#these correspond to c(20,50,100,200)
    paste(samp_size[i])

    bootsamples <- bootsampling(x = sample_list_part1[[i]],
                                boot.replicates = bootstrap_replicates[B])
    corr_samp_dist <- matrix(nrow = bootstrap_replicates[B])

    for(j in 1:bootstrap_replicates[B]){
      corr_samp_dist[j] <- cor(bootsamples[,,j][,1], bootsamples[,,j][,2] )
    }

    hist(corr_samp_dist, breaks = 30, freq = FALSE,
         main = paste("Bootstrap sample size =",
                      c(20,50,100,200)[i]), xlab = "Sample Correlations")
  }
}
```

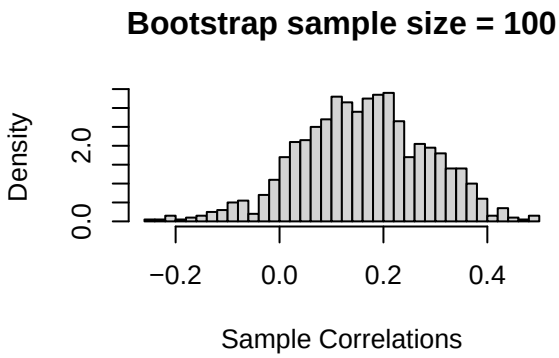
```
## [1] "Bootstrapping with B=  200"
```

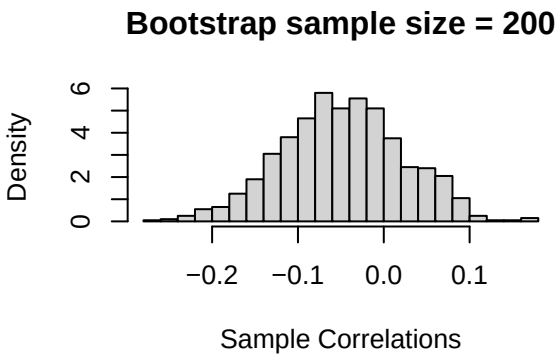


```
## [1] "Bootstrapping with B=  1000"
```

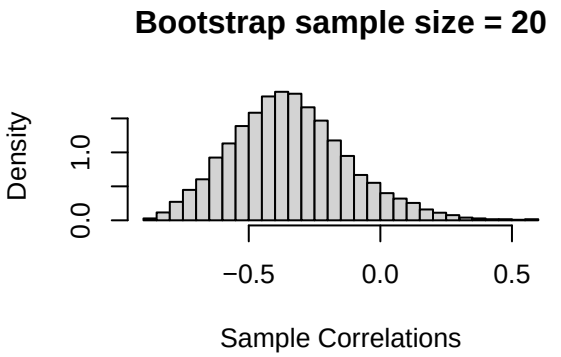


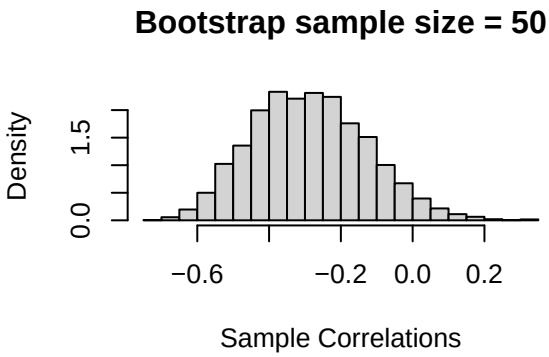


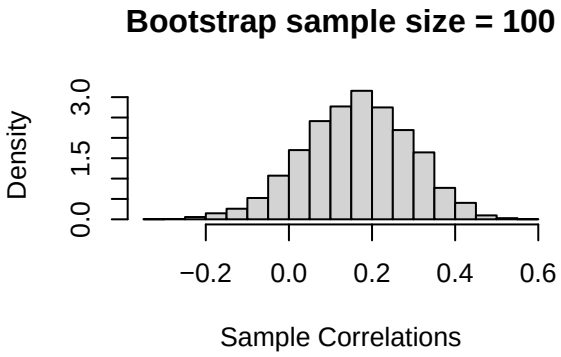


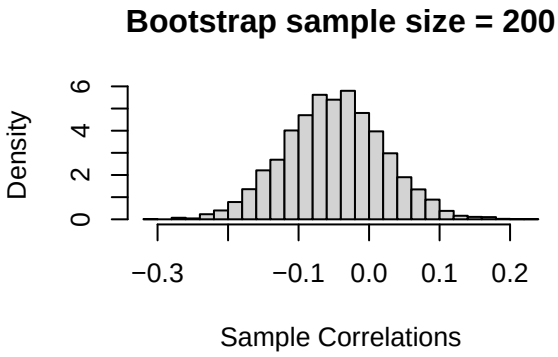



```
## [1] "Bootstrapping with B=  5000"
```

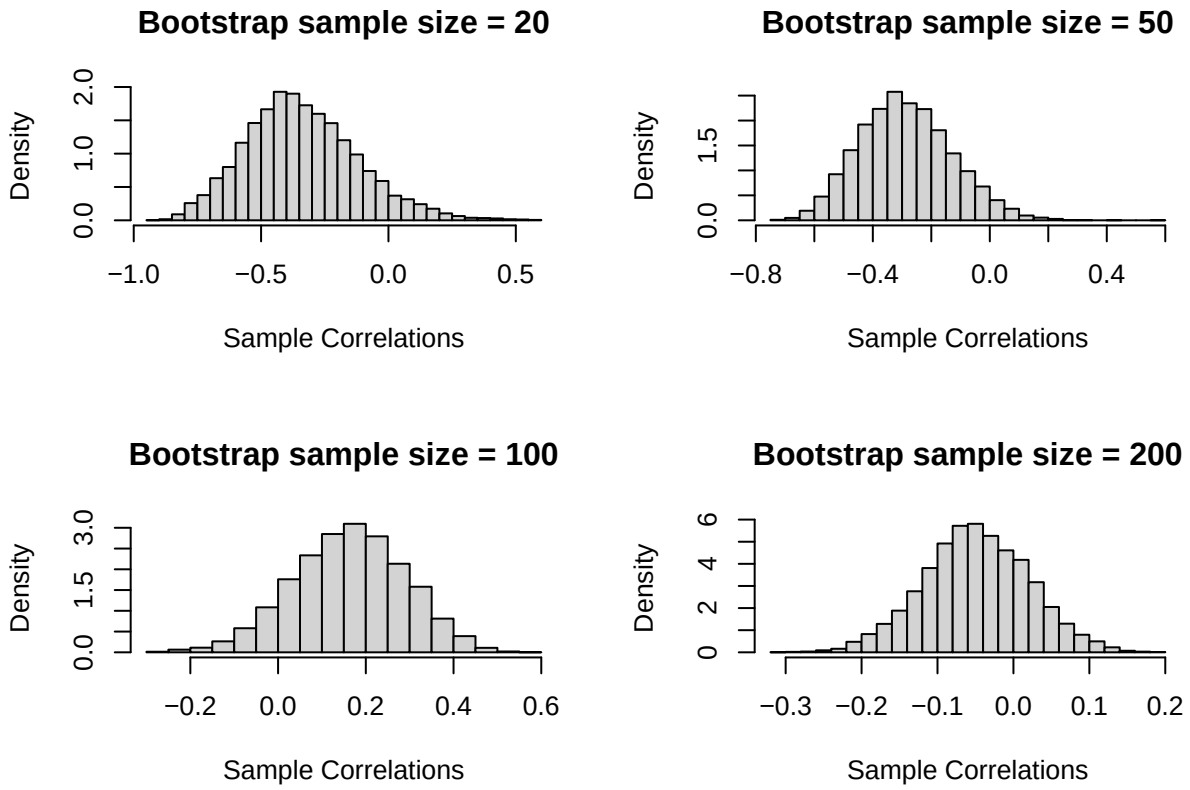








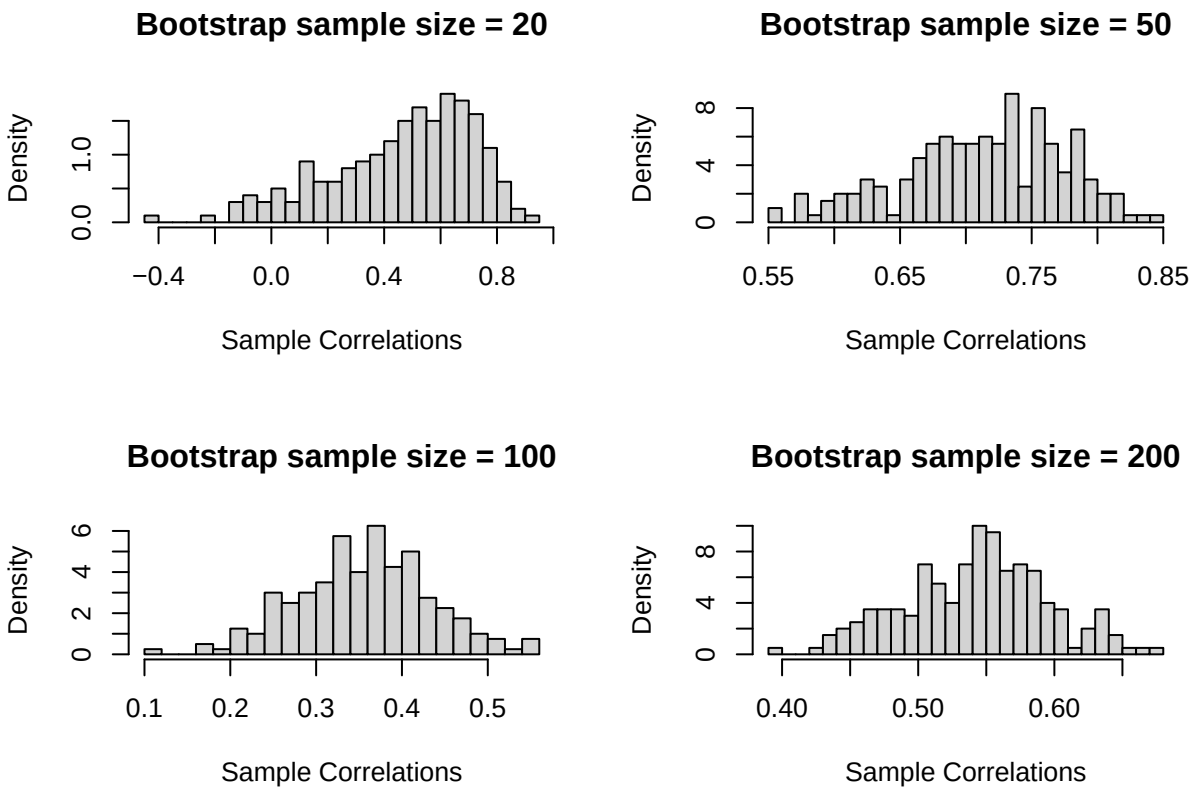

```
## [1] "Bootstrapping with B=  10000"
```



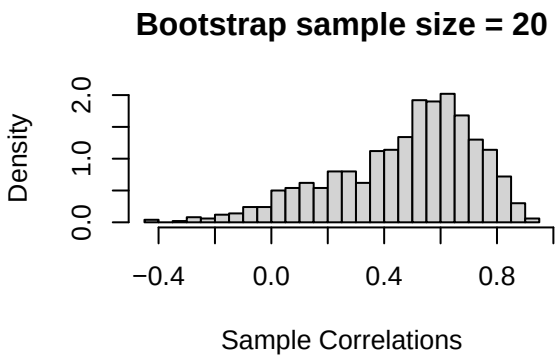
Case 2

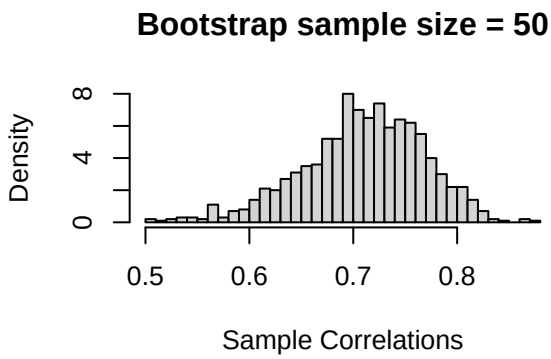
```
set.seed(100)
for(B in seq_along(bootstrap_replicates)){
  print(paste("Bootstrapping with B= ", bootstrap_replicates[B]))
  par(mfrow = c(2,2))
  for(i in seq_along(samp_size)){#these correspond to c(20,50,100,200)
    paste(samp_size[i])

    bootsamples <- bootsampling(x = sample_list_part2[[i]],
                                boot.replicates = bootstrap_replicates[B])
    corr_samp_dist <- matrix(nrow = bootstrap_replicates[B])

    for(j in 1:bootstrap_replicates[B]){
      corr_samp_dist[j] <- cor(bootsamples[,,j][,1], bootsamples[,,j][,2] )
    }

    hist(corr_samp_dist, breaks = 30, freq = FALSE,
         main = paste("Bootstrap sample size =",
                      c(20,50,100,200)[i]), xlab = "Sample Correlations")
  }
}
```

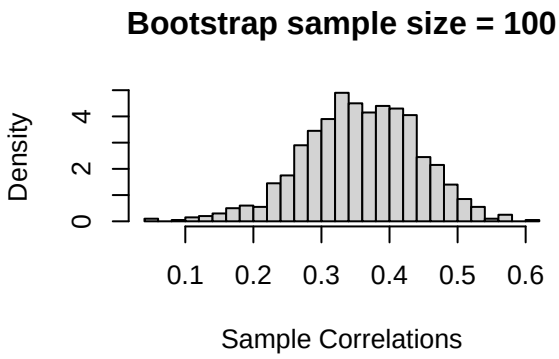
```
## [1] "Bootstrapping with B=  200"
```

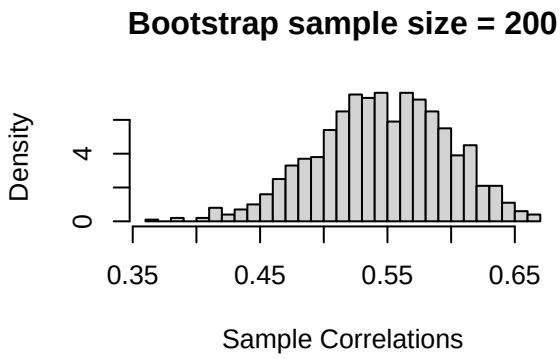



```
## [1] "Bootstrapping with B=  1000"
```

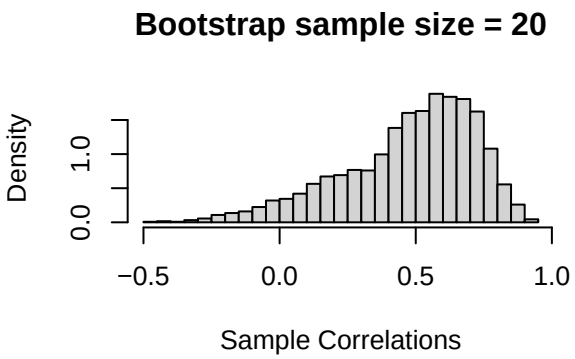


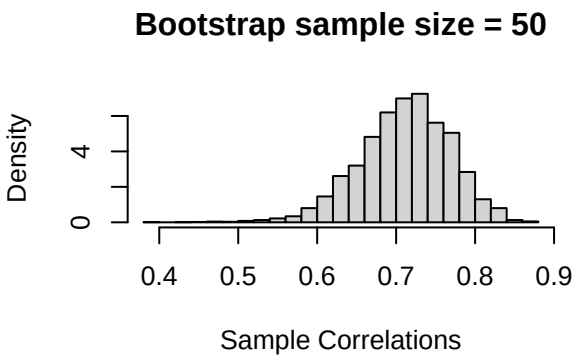


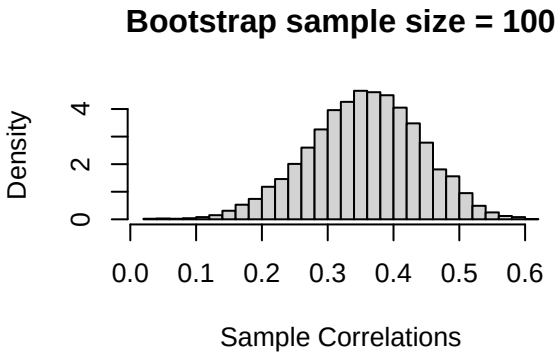


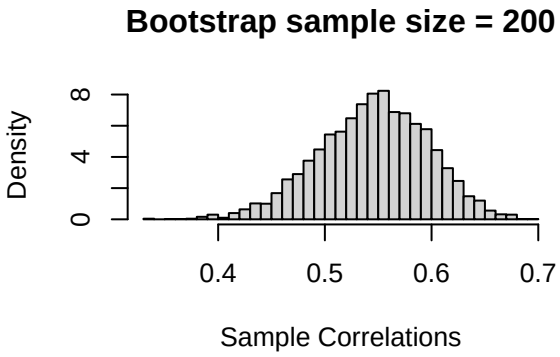



```
## [1] "Bootstrapping with B=  5000"
```

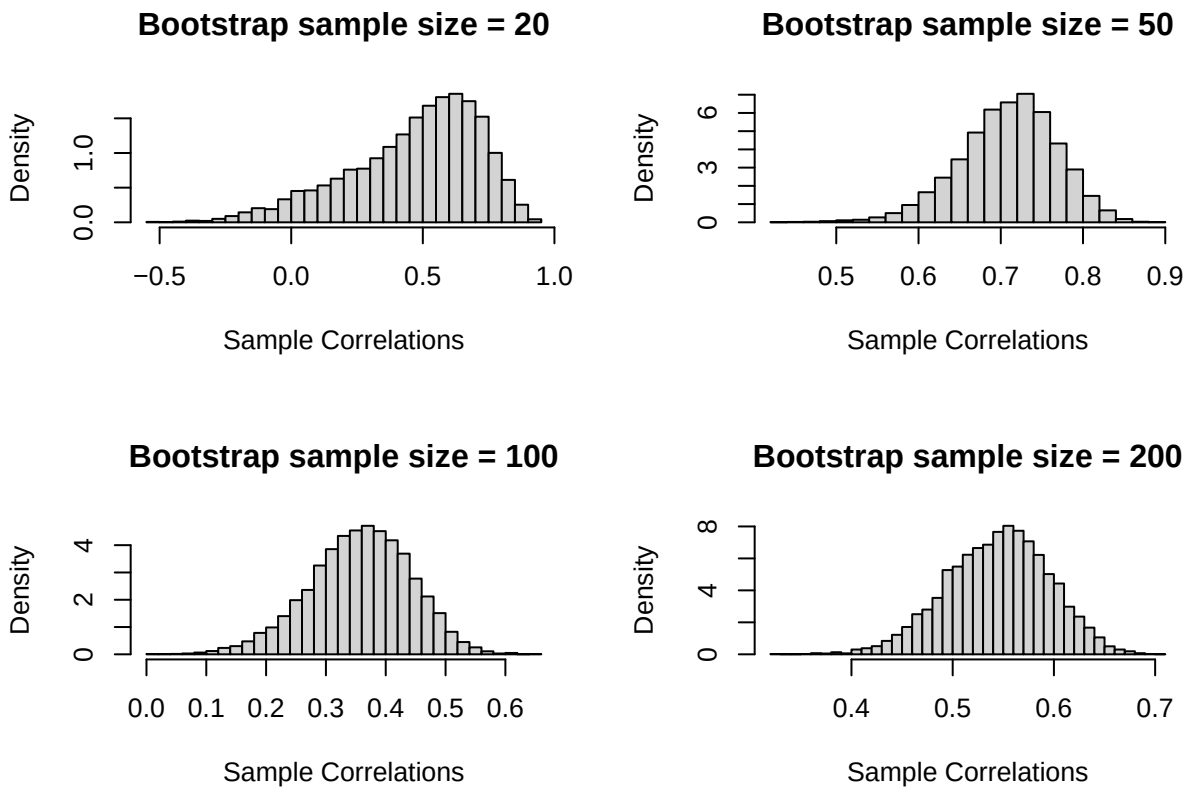








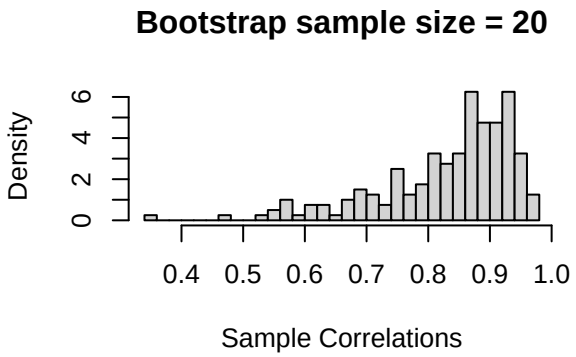

```
## [1] "Bootstrapping with B=  10000"
```

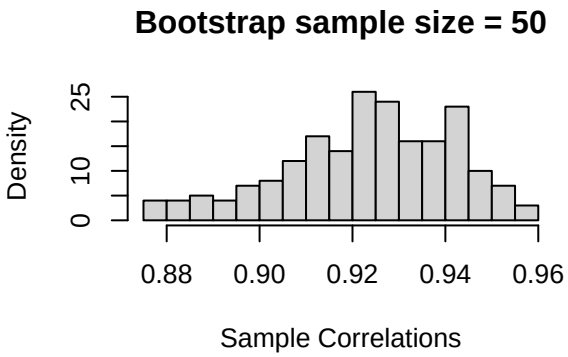


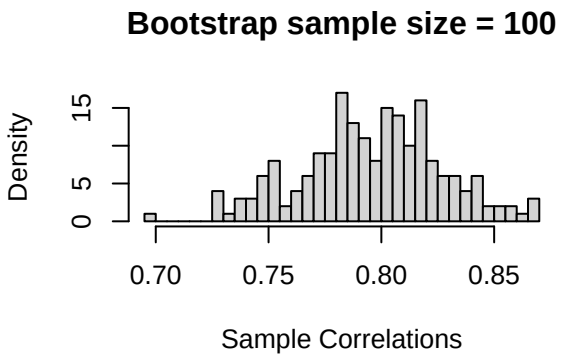
Case 3

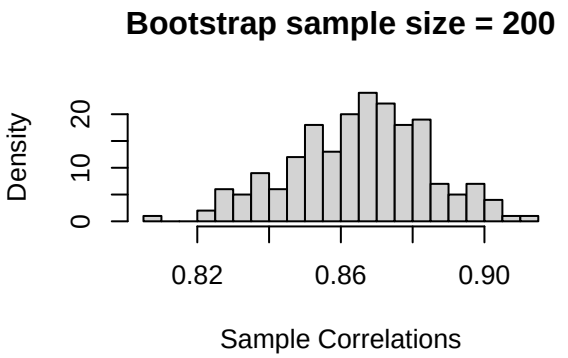
```
set.seed(100)
for(B in seq_along(bootstrap_replicates)){
  print(paste("Bootstrapping with B= ", bootstrap_replicates[B]))
  par(mfrow = c(2,2))
  for(i in seq_along(samp_size)){#these correspond to c(20,50,100,200)
    paste(samp_size[i])

    bootsamples <- bootsampling(x = sample_list_part3[[i]],
                                boot.replicates = bootstrap_replicates[B])
    corr_samp_dist <- matrix(nrow = bootstrap_replicates[B])

    for(j in 1:bootstrap_replicates[B]){
      corr_samp_dist[j] <- cor(bootsamples[,,j][,1], bootsamples[,,j][,2] )
    }

    hist(corr_samp_dist, breaks = 30, freq = FALSE,
         main = paste("Bootstrap sample size =",
                      c(20,50,100,200)[i]), xlab = "Sample Correlations")
  }
}
```

```
## [1] "Bootstrapping with B=  200"
```

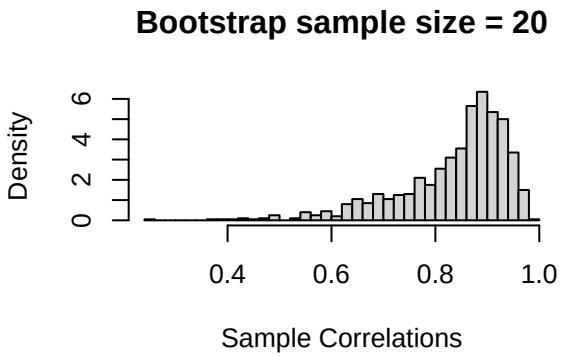


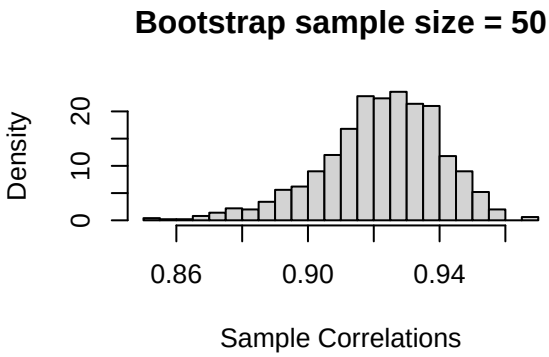


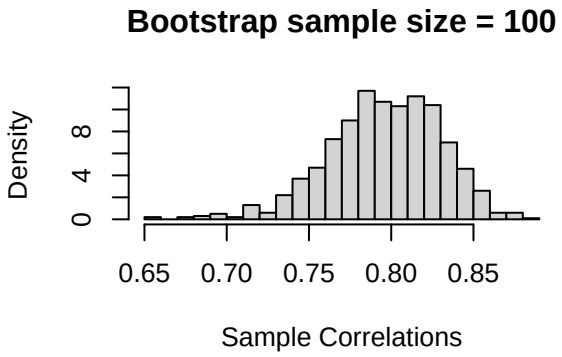


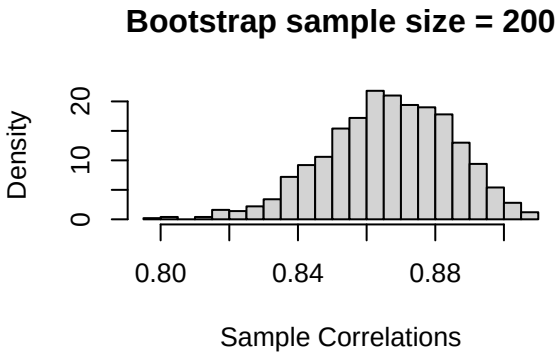



```
## [1] "Bootstrapping with B=  1000"
```

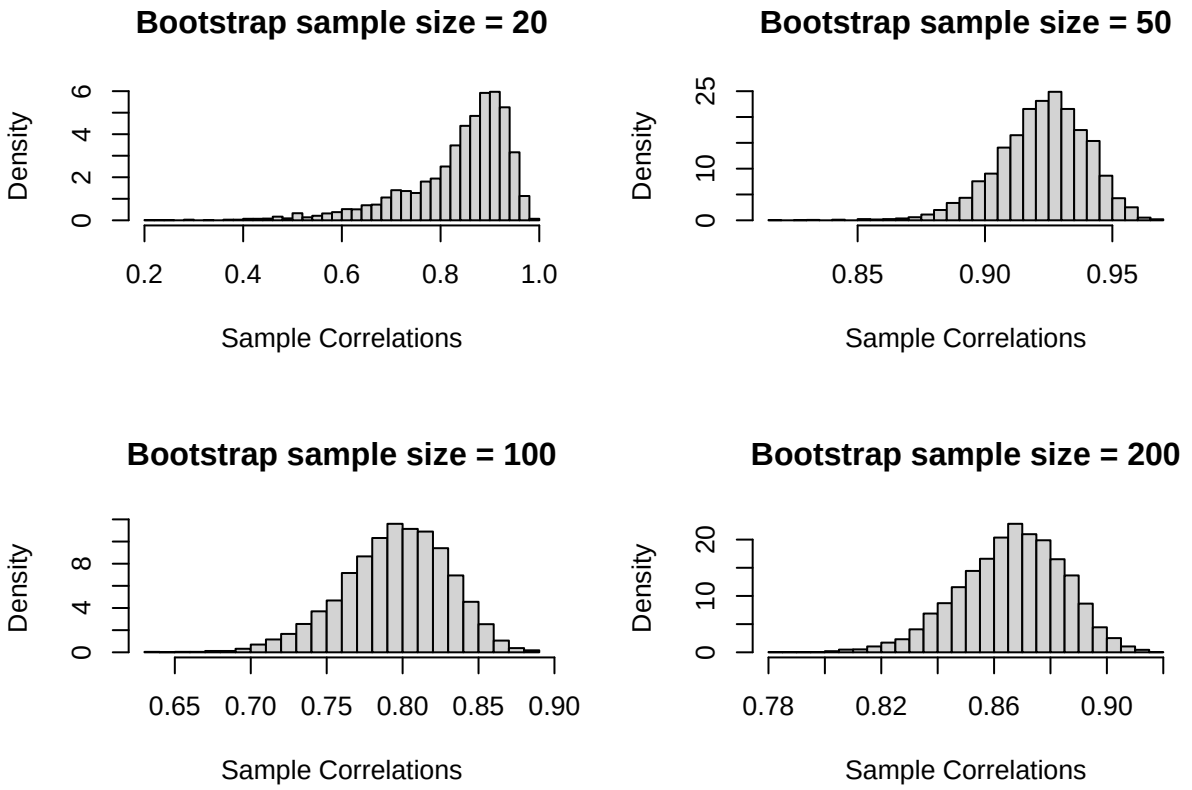




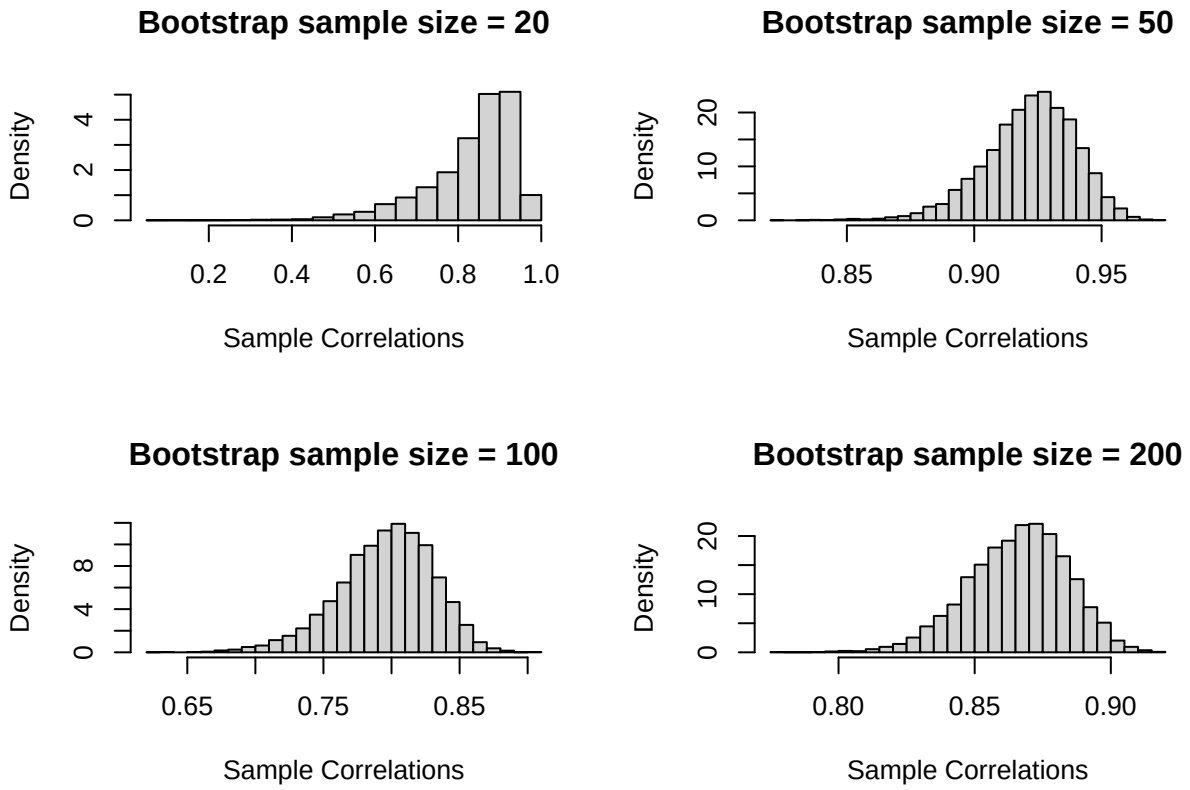





```
## [1] "Bootstrapping with B=  5000"
```




```
## [1] "Bootstrapping with B=  10000"
```



Result Comments:

- As B gets larger, the sampling distribution looks more and more smooth- follows a normal distribution
- For large n, the sample size is larger thus can better represent the true distribution
- When our initial sample is skewed and does not represent the true distribution, the bootstrapping emphasizes this and our sampling distribution is centered around the wrong values.

Problem 2

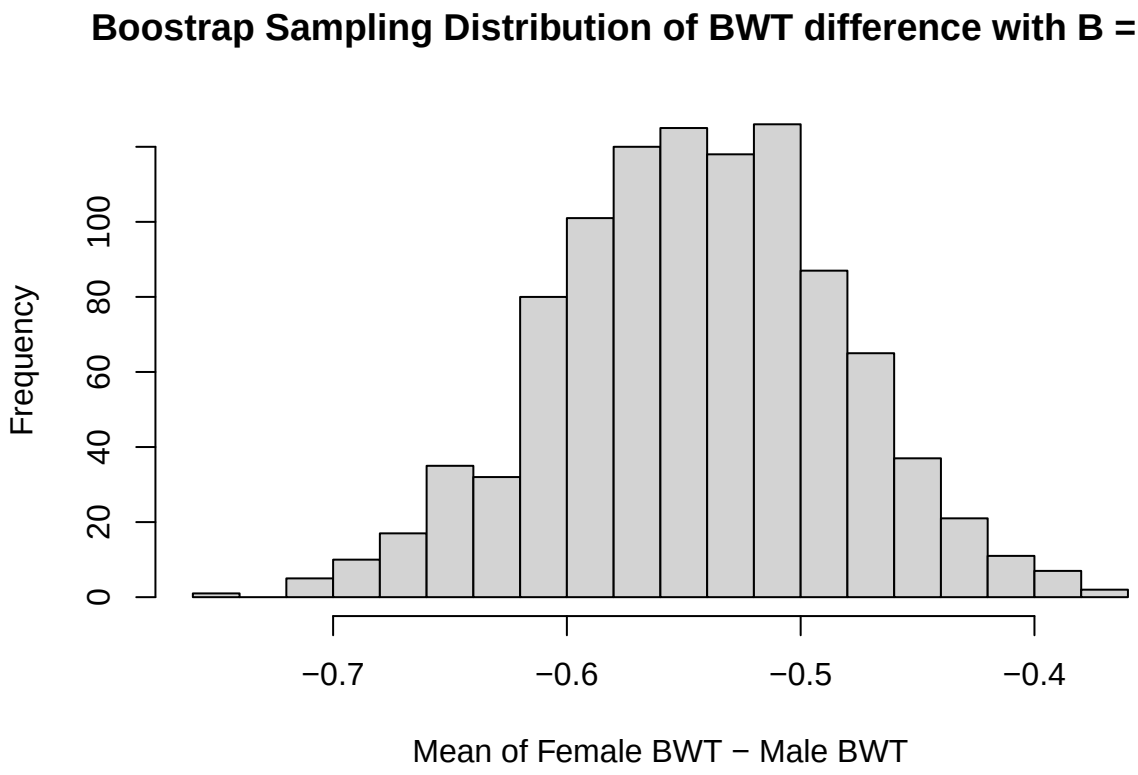
```
data(cats)
summary(cats)
```

```
##  Sex         Bwt             Hwt
##  F:47   Min.   :2.000   Min.   : 6.30
##  M:97   1st Qu.:2.300   1st Qu.: 8.95
##         Median :2.700   Median :10.10
##         Mean   :2.724   Mean   :10.63
##         3rd Qu.:3.025   3rd Qu.:12.12
##         Max.   :3.900   Max.   :20.50
```

part a

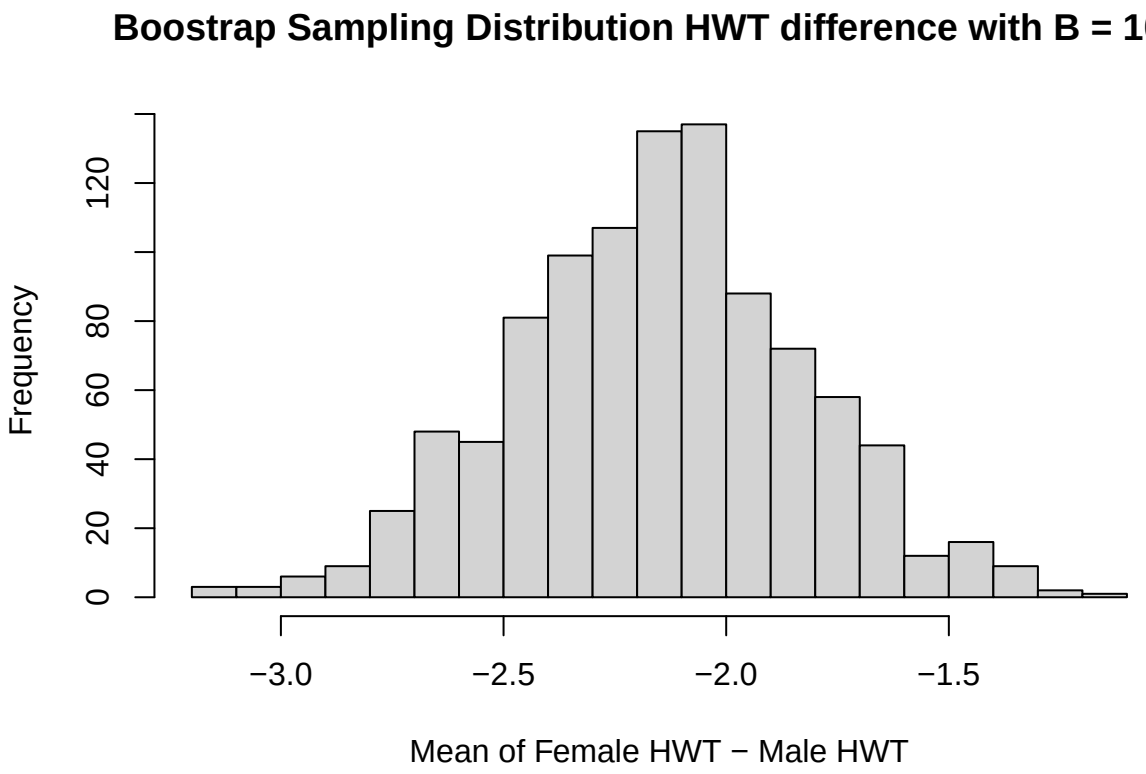
Obtain the bootstrap sampling distribution of the difference of sample means for body weight between female and male cats

```r
set.seed(1)
boot.sample_F <- bootsampling(cats$Bwt[cats$Sex == "F"], boot.replicates = 1000)
sampling_dist_F <- apply(boot.sample_F,2,  mean)
boot.sample_M <- bootsampling(cats$Bwt[cats$Sex == "M"], boot.replicates = 1000)
sampling_dist_M <- apply(boot.sample_M,2,  mean)
hist(sampling_dist_F- sampling_dist_M, breaks = 20, main = paste("Boostrap Sampling Distribution of BWT
```



Obtain the bootstrap sampling distribution of the difference of sample means for heart weight between female and male cats

```r
set.seed(1)
boot.sample_F <- bootsampling(cats$Hwt[cats$Sex == "F"], boot.replicates = 1000)
sampling_dist_F <- apply(boot.sample_F,2,  mean)
boot.sample_M <- bootsampling(cats$Hwt[cats$Sex == "M"], boot.replicates = 1000)
sampling_dist_M <- apply(boot.sample_M,2,  mean)
hist(sampling_dist_F- sampling_dist_M, breaks = 20, main = paste("Boostrap Sampling Distribution HWT di
```



Explain how many bootstrap replicates you decided to use and comment on the results.

I chose to use 1000 bootstraps. I did this because the mean and difference in means, is not an extreme statistic and the Bootstrappign sample shows that 1000 replicates result in a normal sampling distribution.

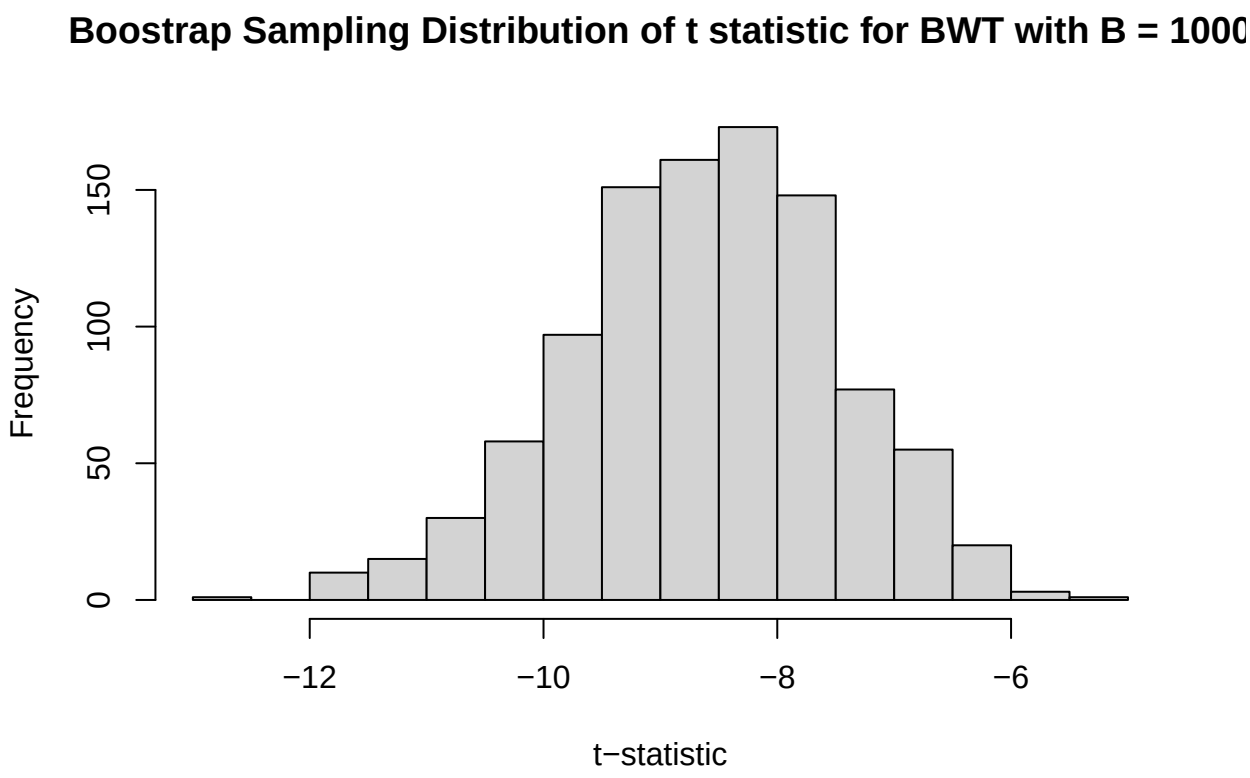
Based on the results, the sampling distributions show that the mean Body Weight of Female Cats is lower Male Cats and the mean Heart Weight of Female Cats is lower Male Cats. We conclude this because the sampling distribution of the sample statistic: mean(Female Weight)- mean(Male Weight), is centered around -0.55 for Body Weight, and -2.2 for Heart Weight. The sampling distribution follows a normal distribution curve where the 95% confidence intervals do not include the value 0, therefore there is significant evidence that the Weight of Female Cats Body weight and Heart Weight is lower than Male Cats.

Part b

Obtain the bootstrap sampling distribution of the t-statistic when testing for mean differences for body weight between female and male cats

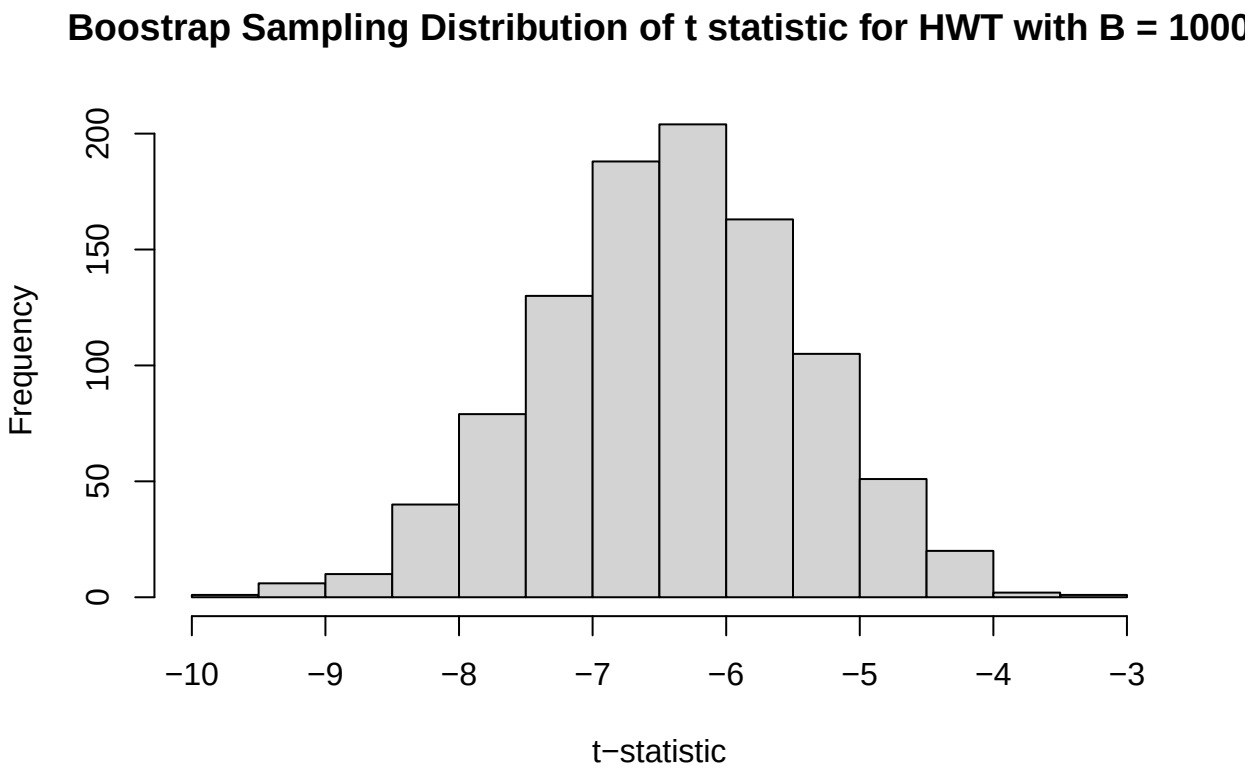
```
set.seed(1)
boot.sample_F <- bootsampling(cats$Bwt[cats$Sex == "F"], boot.replicates = 1000)
sampling_dist_F <- apply(boot.sample_F,2,  mean)

boot.sample_M <- bootsampling(cats$Bwt[cats$Sex == "M"], boot.replicates = 1000)
sampling_dist_M <- apply(boot.sample_M,2,  mean)

diff_means <- sampling_dist_F-sampling_dist_M

denom_values <- sqrt((apply(boot.sample_F,2,  var) + apply(boot.sample_M,2,  var))/2)* sqrt(2/(144/2))
```

```r
t_statistics_boot <- (diff_means/denom_values)

hist(t_statistics_boot, breaks = 20, main = paste("Boostrap Sampling Distribution of t statistic for BW
```



Obtain the bootstrap sampling distribution of the t-statistics when testing for mean differences for heart weight between female and male cats

```r
set.seed(1)
boot.sample_F <- bootsampling(cats$Hwt[cats$Sex == "F"], boot.replicates = 1000)
sampling_dist_F <- apply(boot.sample_F,2,  mean)

boot.sample_M <- bootsampling(cats$Hwt[cats$Sex == "M"], boot.replicates = 1000)
sampling_dist_M <- apply(boot.sample_M,2,  mean)

diff_means <- sampling_dist_F-sampling_dist_M

denom_values <- sqrt((apply(boot.sample_F,2,  var) + apply(boot.sample_M,2,  var))/2)* sqrt(2/(144/2))


t_statistics_boot <- (diff_means/denom_values)

hist(t_statistics_boot, breaks = 20, main = paste("Boostrap Sampling Distribution of t statistic for HW
```



Explain how many bootstrap replicates you decided to use and comment on the results.

I chose to use 1000 bootstraps. This is because like the difference in means, the t-statistic is not an extreme value and using 1000 bootstraps, the resulting sampling distribution shows a normal distribution.

From the results, the t statistic for the difference in body weight between males and females is centered around -9, indicating there is a difference in body weight between the two genders. This is the same result we observe in part a.

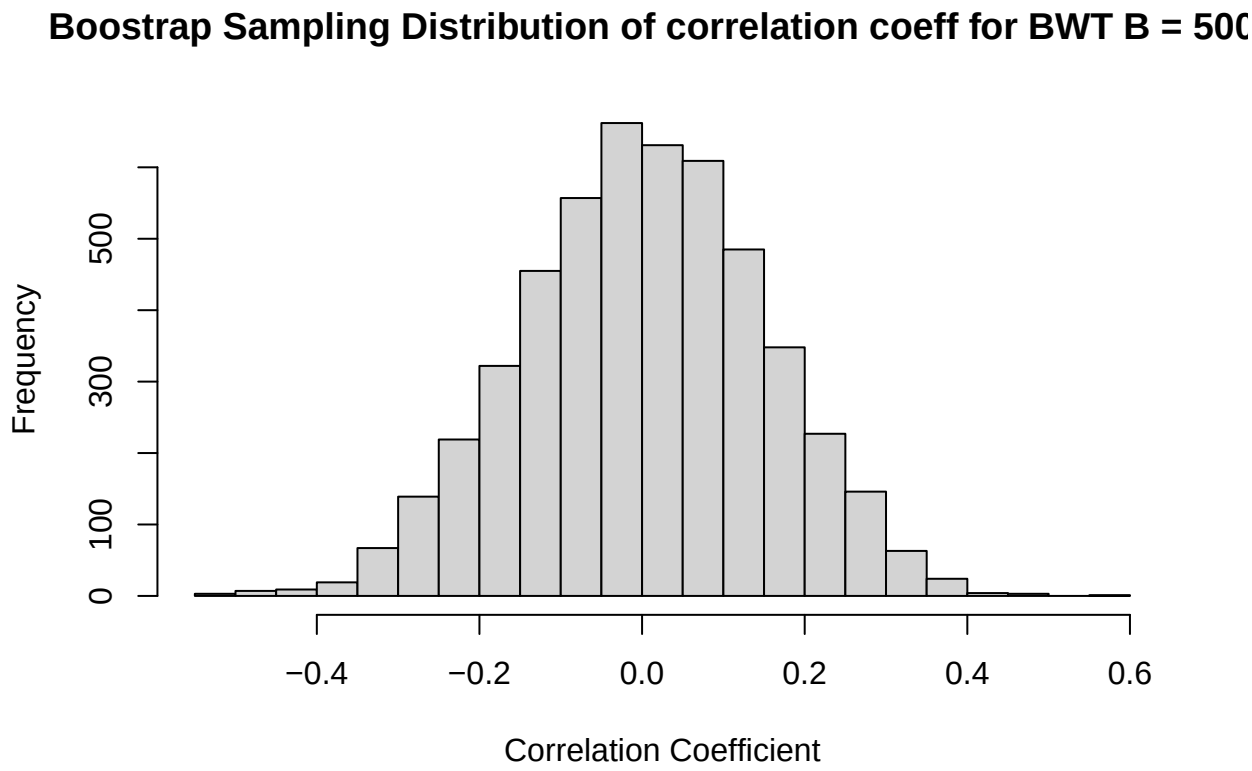
The t-statistic for the difference in heart weight between males and females is centered around -6.5, indicating there is a difference in heart weight between the two genders. Again, this is the same result we observe from part a.

Part c

Using your code from Problem 1(c): - Obtain the bootstrap sampling distribution of the sample correlation coefficient between body weight and heart weight for female cats

```
bootsampling_cats <- function(x, boot.replicates = B){
  x = as.matrix(x)
  nx = 47
  return(replicate(boot.replicates, x[sample.int(nx, replace = TRUE),]))
}


set.seed(1)
boot.sample_F <- bootsampling_cats(cats$Bwt[cats$Sex == "F"], boot.replicates = 5000)
```

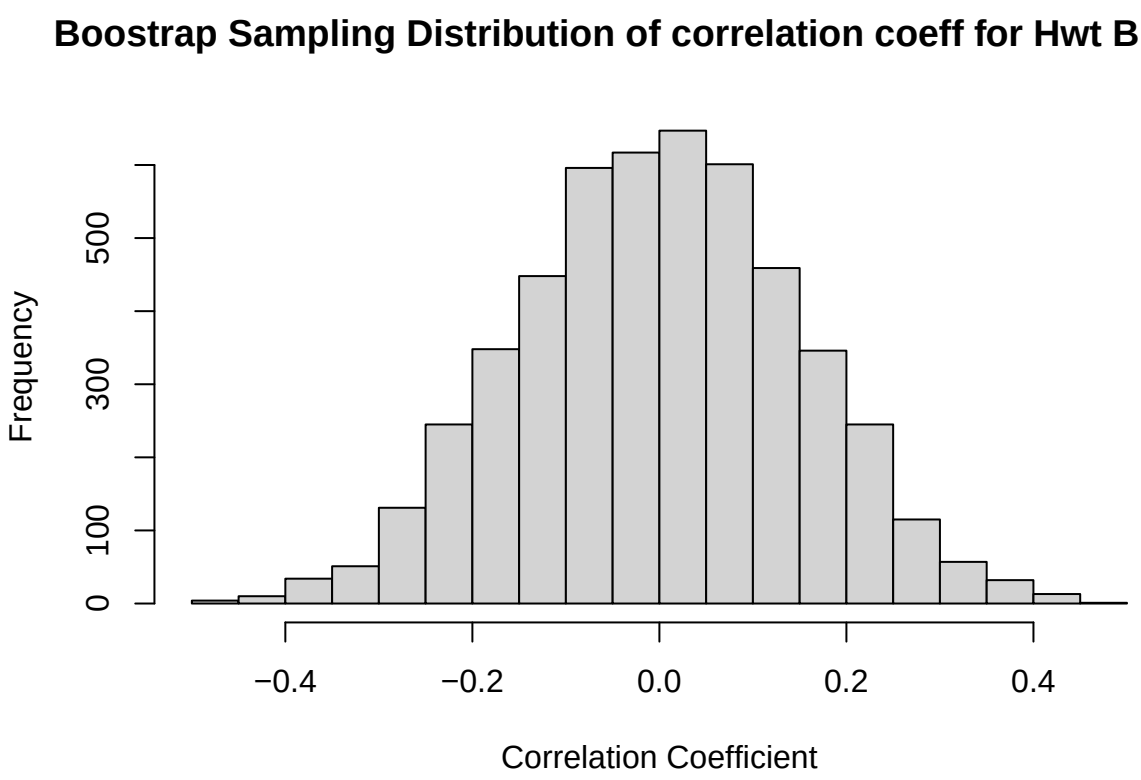
```r
boot.sample_M <- bootsampling_cats(cats$Bwt[cats$Sex == "M"], boot.replicates = 5000)

cor_matrix <- matrix(nrow = 5000)
for(B in 1:5000){
  cor_matrix[B] <- cor(boot.sample_F[,B],boot.sample_M[,B])
}

hist(cor_matrix, breaks = 20, main = paste("Boostrap Sampling Distribution of correlation coeff for BWT
```



- Obtain the bootstrap sampling distribution of the sample correlation coefficient between body weight and heart weight for male cats

```r
bootsampling_cats <- function(x, boot.replicates = B){
  x = as.matrix(x)
  nx = 47
  return(replicate(boot.replicates, x[sample.int(nx, replace = TRUE),]))
}


set.seed(1)
boot.sample_F <- bootsampling_cats(cats$Hwt[cats$Sex == "F"], boot.replicates = 5000)

boot.sample_M <- bootsampling_cats(cats$Hwt[cats$Sex == "M"], boot.replicates = 5000)

cor_matrix <- matrix(nrow = 5000)
for(B in 1:5000){
  cor_matrix[B] <- cor(boot.sample_F[,B],boot.sample_M[,B])
```

```
}

hist(cor_matrix, breaks = 20, main = paste("Boostrap Sampling Distribution of correlation coeff for Hwt
```



Explain how many bootstrap replicates you decided to use and comment on the results.

For finding the sampling distribution of correlation coefficient for body weight between cat weight and male weight, I decided to use 5000 replicates. This is because by using more repetitions, the resulting sampling distribution gets closer to the true distribution, and I was unsure about the true shape of the correlation so I wanted to use more repetitions.

The resulting sampling distribution shows that the mean of correlation coefficients for the bootstrap samples is centered around 0 to 0.05. This indicates that the weights of female cats and male cats are not correlated.