Lab 09

Jennifer Lin

jenniferyjlin@berkeley.edu (mailto:jenniferyjlin@berkeley.edu)

```
library(datasets)
library(boot)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

1

Load the iris dataset, and find the mean and SEM of the Sepal.Length, Sepal.Width, Petal.Length, Petal.Width.

```
data(iris)
variableList <- c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width")
for(i in 1:4){
  vec <- iris[ , variableList[i]]
  vec_mean <- mean(vec)
  vec_SEM <- sd(vec)/sqrt(length(vec))
  print(paste0(variableList[i], ": mean=", round(vec_mean,4), ", SEM=", round(vec_SE
M,4)))
}
```

```
## [1] "Sepal.Length: mean=5.8433, SEM=0.0676"
## [1] "Sepal.Width: mean=3.0573, SEM=0.0356"
## [1] "Petal.Length: mean=3.758, SEM=0.1441"
## [1] "Petal.Width: mean=1.1993, SEM=0.0622"
```

2

Calculate the variance of Sepal.Length, Sepal.Width, Petal.Length, Petal.Width.

```
for(i in 1:4){
  vec <- iris[ , variableList[i]]
  vec_variance <- var(vec)
  print(paste0(variableList[i], ": variance=", round(vec_variance,4)))
}
```

```
## [1] "Sepal.Length: variance=0.6857"
## [1] "Sepal.Width: variance=0.19"
## [1] "Petal.Length: variance=3.1163"
## [1] "Petal.Width: variance=0.581"
```

3

Using the values calculated above, obtain the 95% confidence interval of the mean for Sepal.Length, Sepal.Width, Petal.Length, Petal.Width.

What determines the range of the confidence interval?

```
for(i in 1:4){
  vec <- iris[ , variableList[i]]
  vec_SEM <- sd(vec)/sqrt(length(vec))
  # 95% CI using normal distribution
  CI_n_lower <- mean(vec)-qnorm(0.975)*vec_SEM
  CI_n_upper <- mean(vec)+qnorm(0.975)*vec_SEM
  # 95% CI using t distribution
  CI_t_lower <- mean(vec)-qt(0.975,df=length(vec)-1)*vec_SEM
  CI_t_upper <- mean(vec)+qt(0.975,df=length(vec)-1)*vec_SEM
  print(paste0(variableList[i], ": 95% CI (using normal distribution)=", round(CI_n_l
ower,4), "~",round(CI_n_upper,4), ", 95% CI (using t distribution)=", round(CI_t_lowe
r,4), "~",round(CI_t_upper,4)))
}
```

```
## [1] "Sepal.Length: 95% CI (using normal distribution)=5.7108~5.9758, 95% CI (using
t distribution)=5.7097~5.9769"
## [1] "Sepal.Width: 95% CI (using normal distribution)=2.9876~3.1271, 95% CI (using
t distribution)=2.987~3.1277"
## [1] "Petal.Length: 95% CI (using normal distribution)=3.4755~4.0405, 95% CI (using
t distribution)=3.4732~4.0428"
## [1] "Petal.Width: 95% CI (using normal distribution)=1.0774~1.3213, 95% CI (using
t distribution)=1.0764~1.3223"
```

4

Obtain 500 bootstrap samples out of Sepal.Length, Sepal.Width, Petal.Length, Petal.Width:

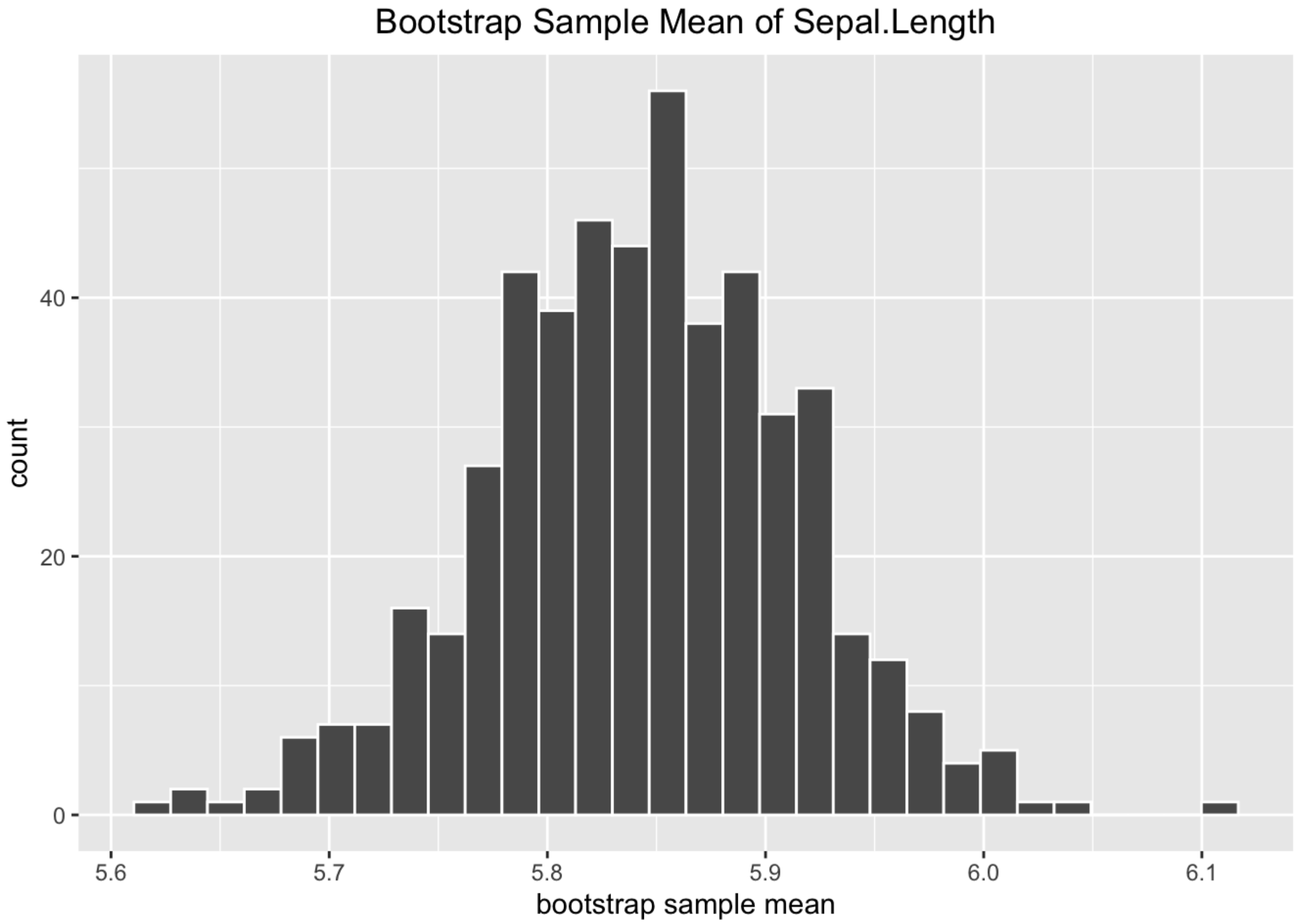
(a) Create a histogram of the bootstrap sample means

(b) Find the 95% bootstrap confidence intervals

(c) Write a function for your own estimator (something other than the mean - could be for instance the median or skew) and compute the 95% confidence interval for it.

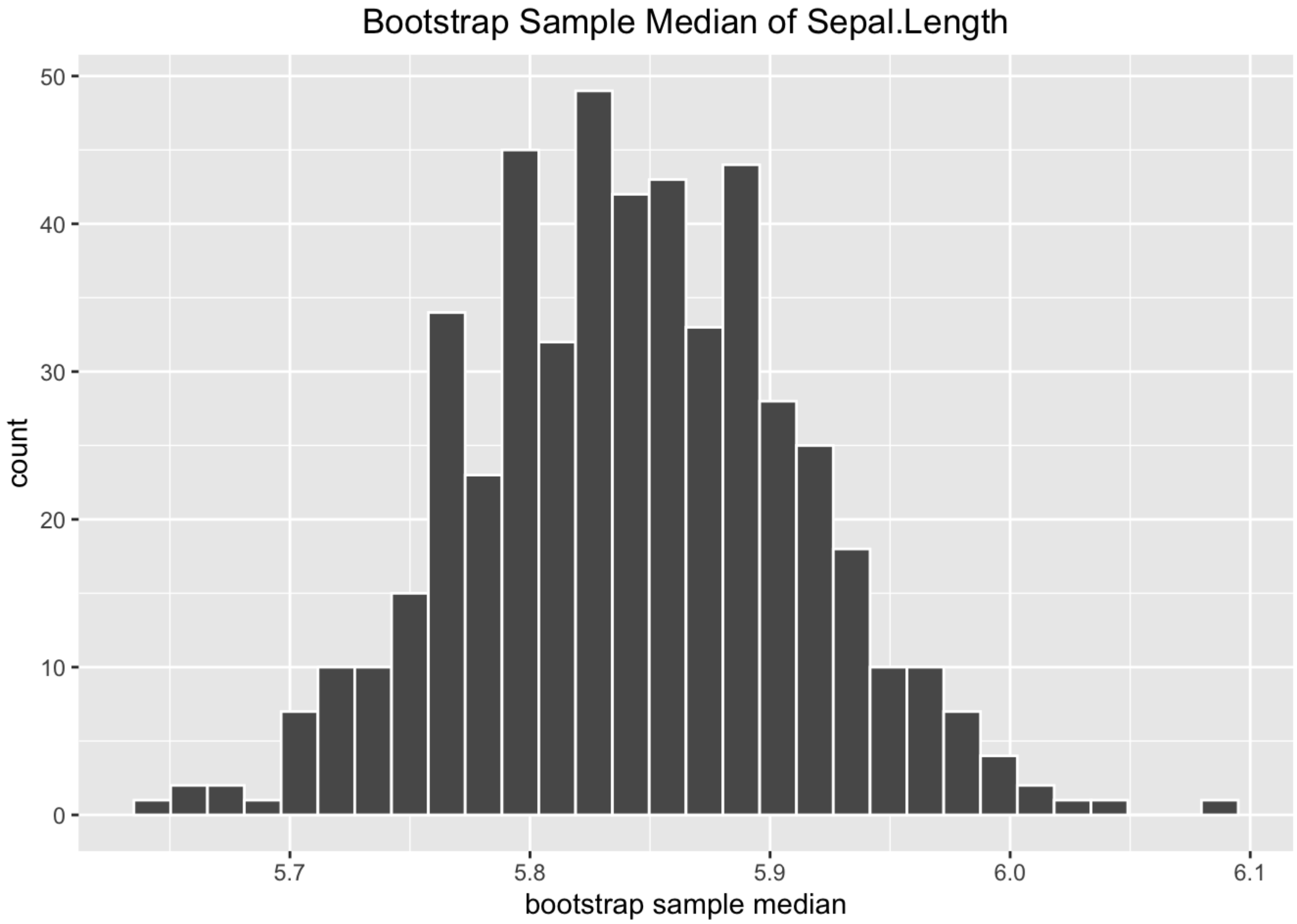
```
for(i in 1:4){
  vec <- iris[ , variableList[i]]
  ### BOOTSTRAP ##############################################
  vec_bt_mean <- c()
  vec_bt_median <- c()
  for (j in 1:500){
    vec_bt_mean <- c(vec_bt_mean, mean(sample(vec,length(vec),replace=T)))
    vec_bt_median <- c(vec_bt_median, mean(sample(vec,length(vec),replace=T)))
  }
  ### HISTOGRAM ##############################################
  # mean
  p_CI_bt_mean <- ggplot() + aes(vec_bt_mean) + geom_histogram(colour="white") +
    labs(title=paste0("Bootstrap Sample Mean of ", variableList[i])) +
    xlab("bootstrap sample mean") +
    theme(plot.title = element_text(hjust = 0.5))
  # median
  p_CI_bt_median <- ggplot() + aes(vec_bt_median) + geom_histogram(colour="white") +
    labs(title=paste0("Bootstrap Sample Median of ", variableList[i])) +
    xlab("bootstrap sample median") +
    theme(plot.title = element_text(hjust = 0.5))
  ### 95% CI #################################################
  # mean
  CI_bt_mean_lower <- quantile(vec_bt_mean, 0.025)
  CI_bt_mean_upper <- quantile(vec_bt_mean, 0.975)
  # median
  CI_bt_median_lower <- quantile(vec_bt_median, 0.025)
  CI_bt_median_upper <- quantile(vec_bt_median, 0.975)
  ### PRINT RESULTS ##########################################
  print(p_CI_bt_mean)
  print(paste0(variableList[i], ": 95% CI of Mean = ", round(CI_bt_mean_lower,4), "~"
,round(CI_bt_mean_upper,4)))
  print(p_CI_bt_median)
  print(paste0(variableList[i], ": 95% CI of Median = ", round(CI_bt_median_lower,4),
"~",round(CI_bt_median_upper,4)))
}
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



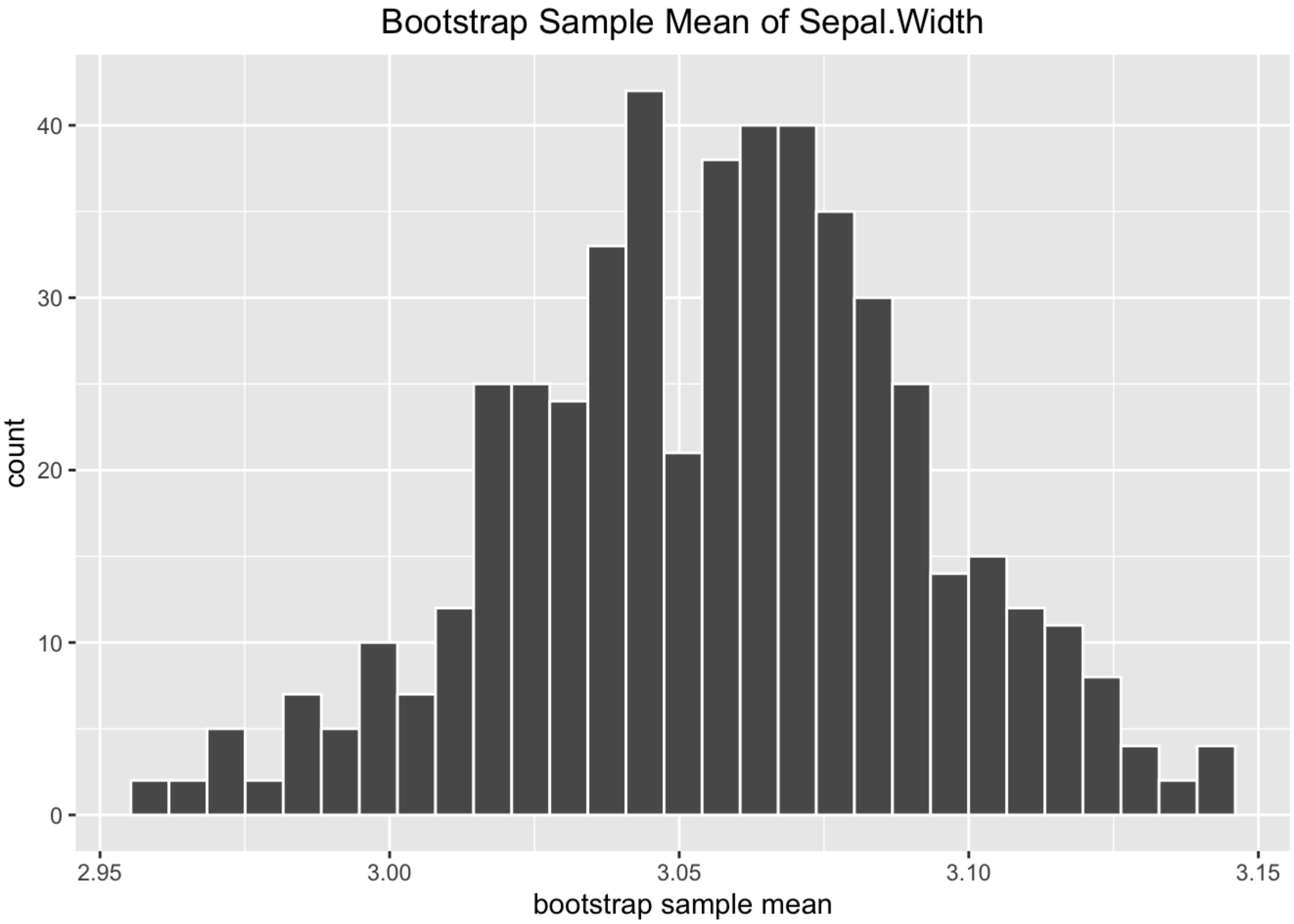
```
## [1] "Sepal.Length: 95% CI of Mean = 5.6979~5.981"
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



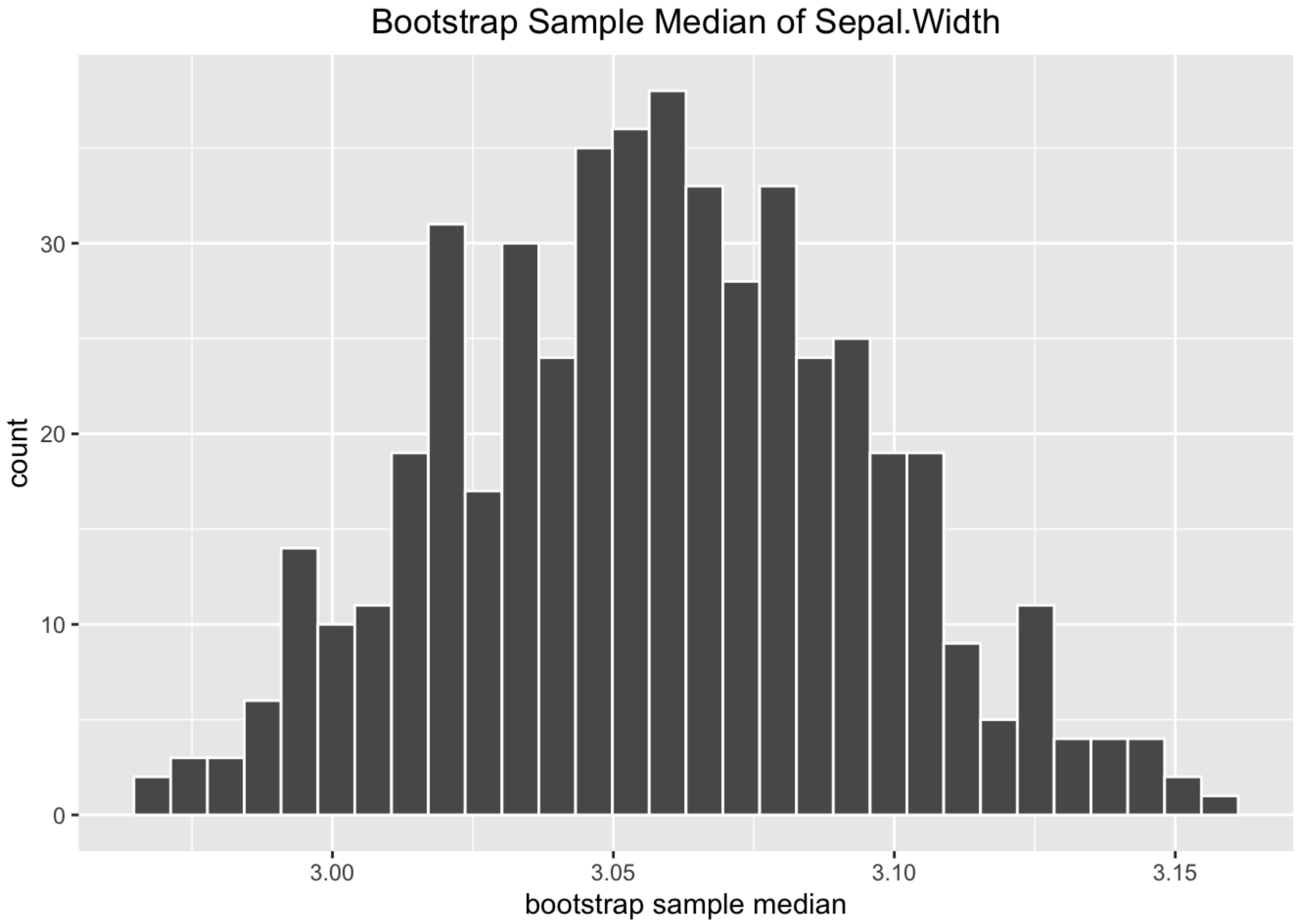
```
## [1] "Sepal.Length: 95% CI of Median = 5.7123~5.9771"
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



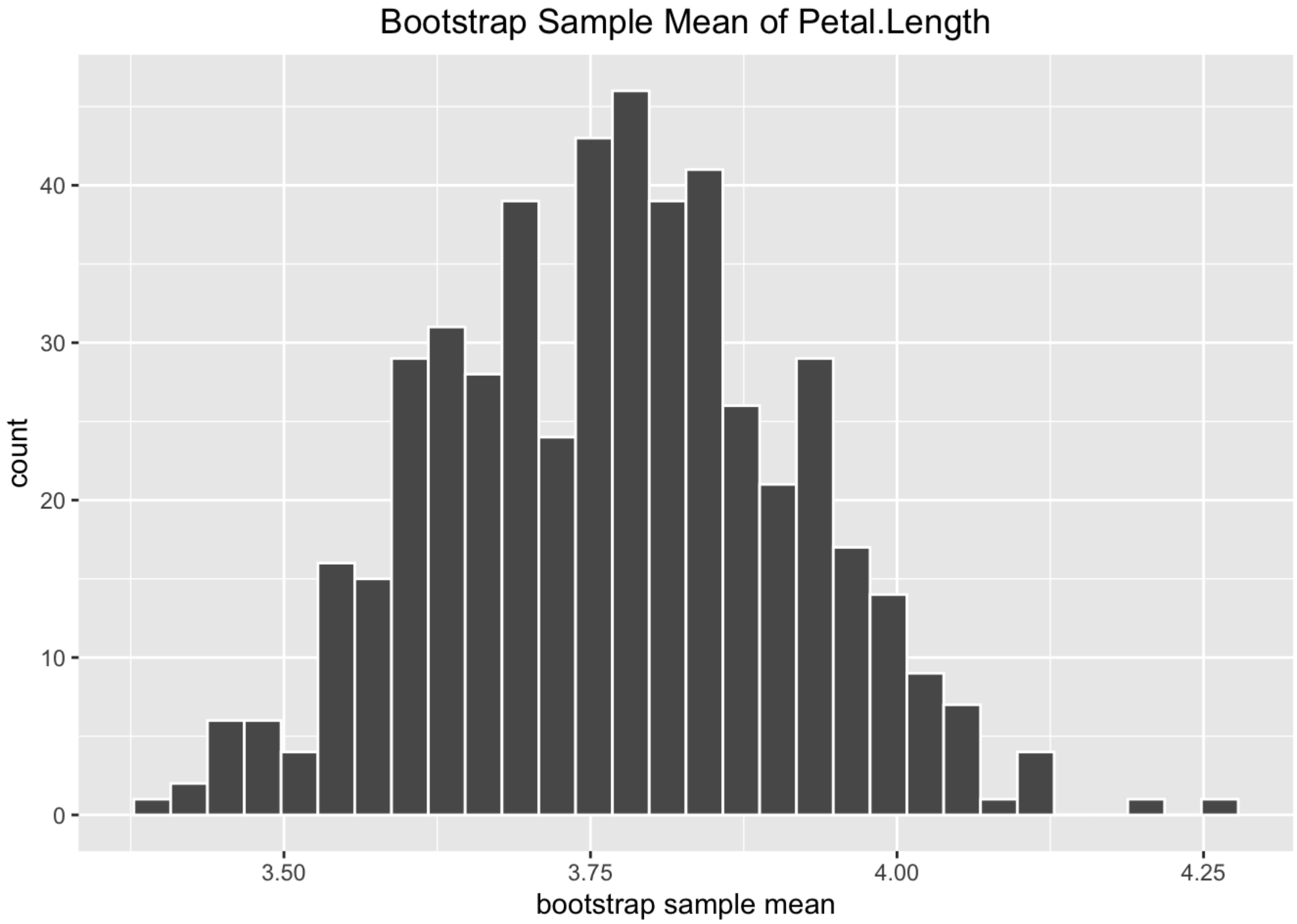
```
## [1] "Sepal.Width: 95% CI of Mean = 2.983~3.123"
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



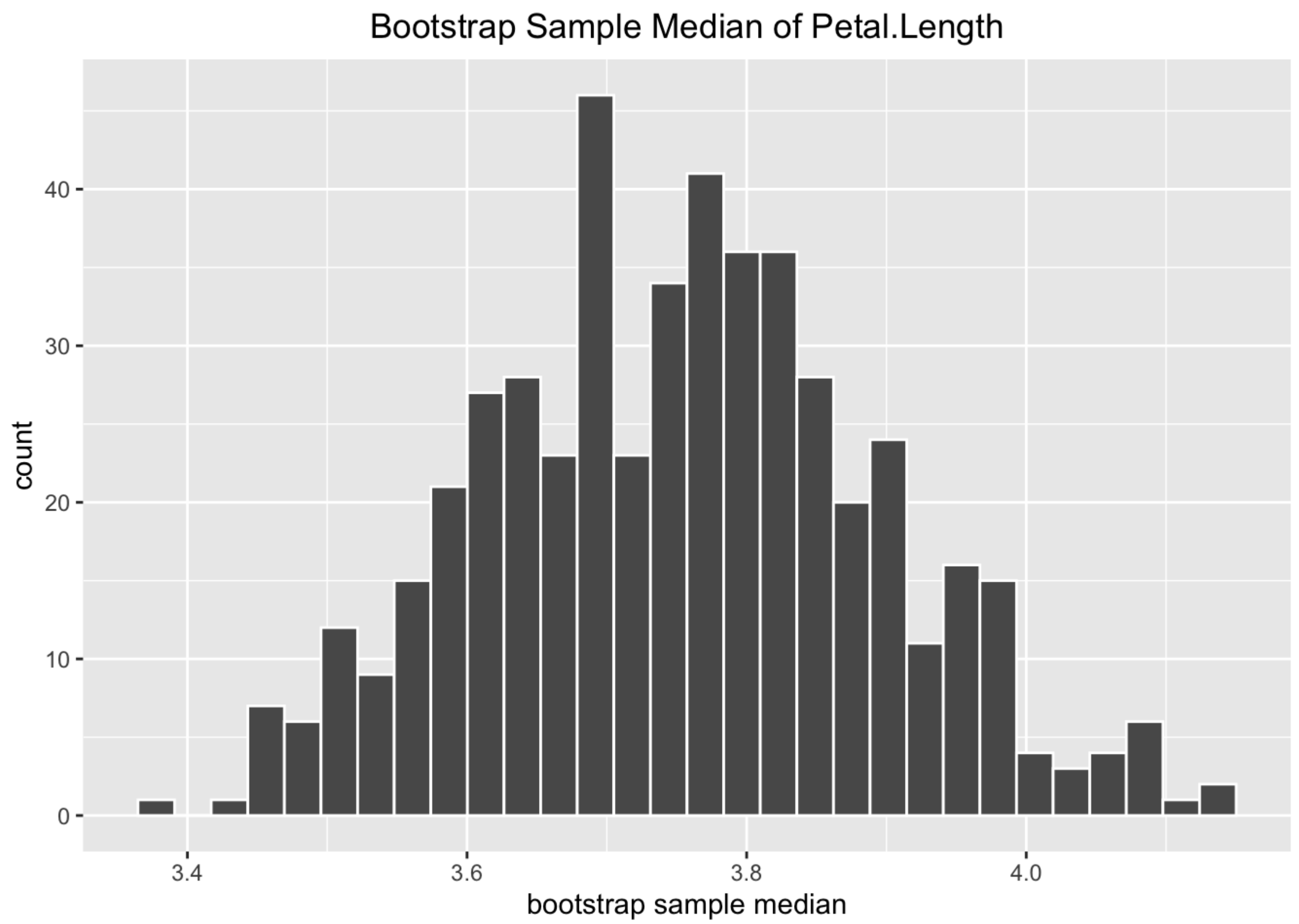
```
## [1] "Sepal.Width: 95% CI of Median = 2.9903~3.1317"
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



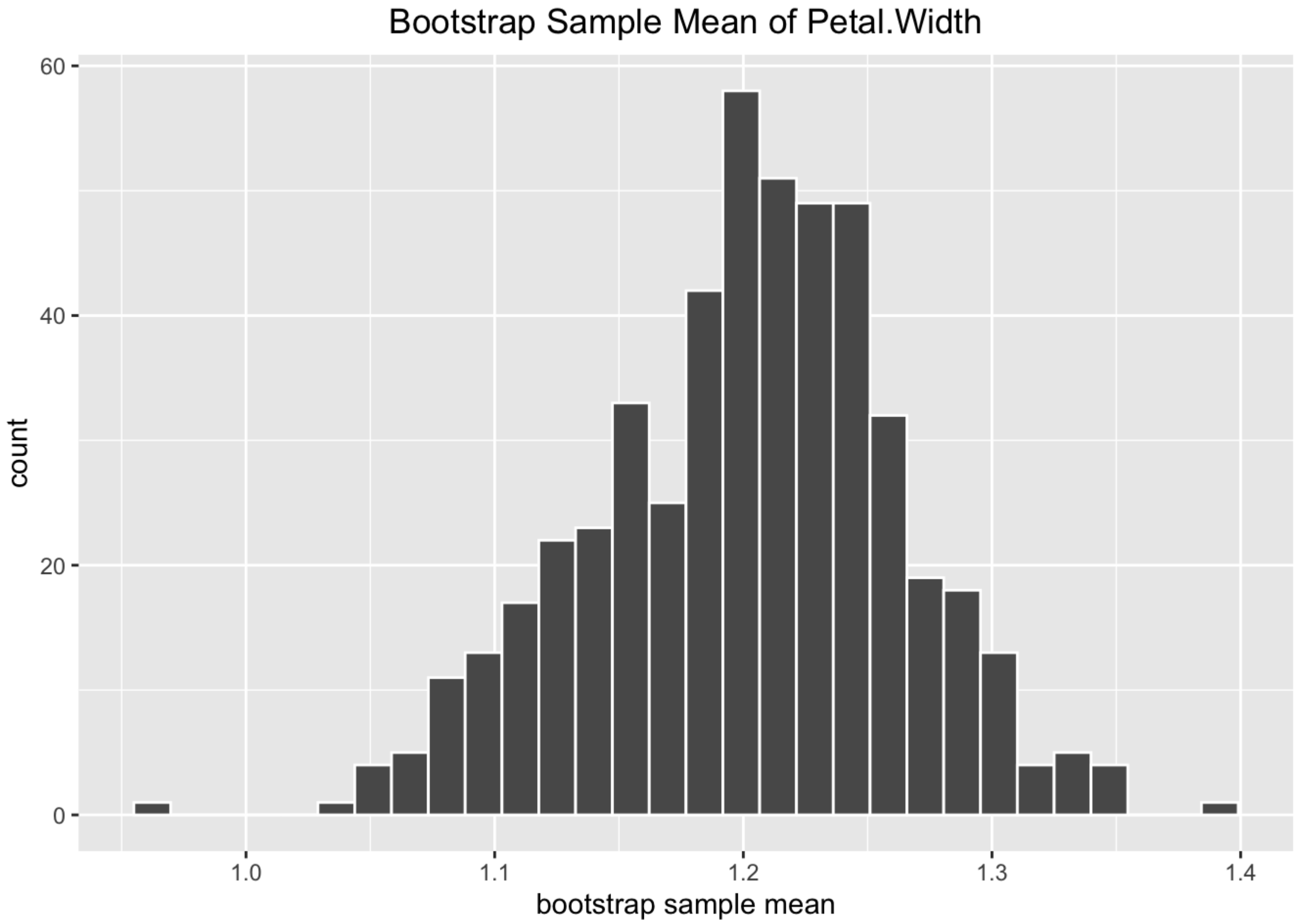
```
## [1] "Petal.Length: 95% CI of Mean = 3.4831~4.0414"
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



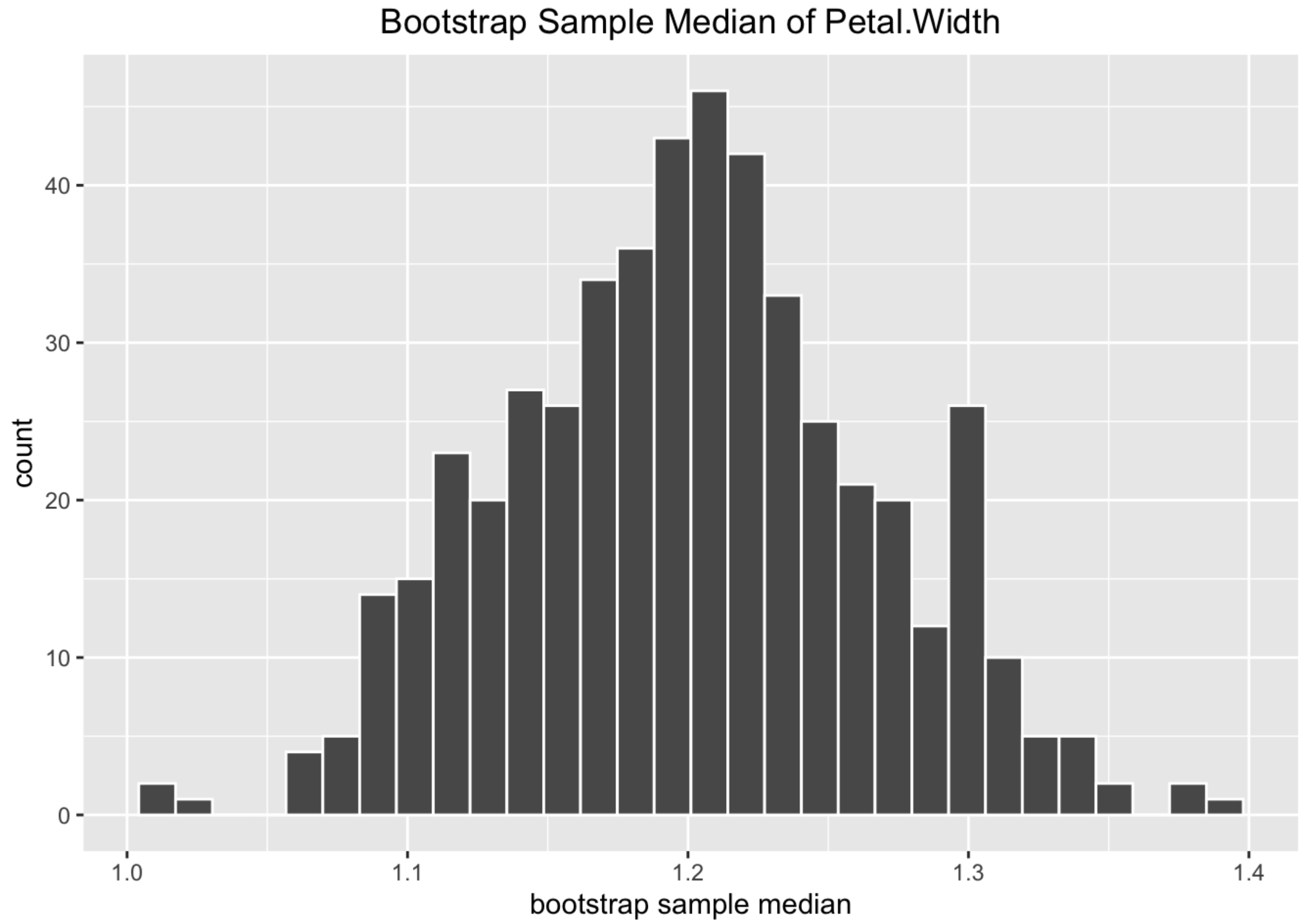
```
## [1] "Petal.Length: 95% CI of Median = 3.4873~4.04"
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
## [1] "Petal.Width: 95% CI of Mean = 1.0776~1.311"
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
## [1] "Petal.Width: 95% CI of Median = 1.0836~1.3233"
```

5

Perform a t-test between all combinations of the four datasets (that would be a total of 4 choose 2 = 6 combinations).

See if there is any statistically significant relationship between any of the four parameters defining the iris.

```
# create all combinations
variableList_combn <- combn(variableList,2)
print(variableList_combn)
```

```
##      [,1]           [,2]           [,3]           [,4]           [,5]
## [1,] "Sepal.Length" "Sepal.Length" "Sepal.Length" "Sepal.Width"  "Sepal.Width"
## [2,] "Sepal.Width"  "Petal.Length" "Petal.Width"  "Petal.Length" "Petal.Width"
##      [,6]
## [1,] "Petal.Length"
## [2,] "Petal.Width"
```

```r
# t-test
for(i in 1:length(variableList_combn[1,])){
  x <- iris[ , variableList_combn[1,i] ]
  y <- iris[ , variableList_combn[2,i] ]
  tTestResult <- t.test(x,y)
  print(paste0("### T-test results of ", variableList_combn[1,i], " and ", variableList_combn[2,i], " ####################################"))
  print(tTestResult)
  if(tTestResult$p.value <= 0.05){
    print(paste0("P value is smaller than (or equal to) 0.05 and thus we can reject the null hypothesis.  Consequently, the means of ", variableList_combn[1,i], " and ", variableList_combn[2,i], " are significantly differnet from each other."))
  }else{
    print(paste0("P value is larger than 0.05 and thus we can NOT reject the null hypothesis.  Consequently, the means of ", variableList_combn[1,i], " and ", variableList_combn[2,i], " are NOT significantly differnet from each other."))
  }
}
```

```
## [1] "### T-test results of Sepal.Length and Sepal.Width ##########################
##########"
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 36.463, df = 225.68, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.63544 2.93656
## sample estimates:
## mean of x mean of y
##  5.843333  3.057333
##
## [1] "P value is smaller than (or equal to) 0.05 and thus we can reject the null hy
pothesis.  Consequently, the means of Sepal.Length and Sepal.Width are significantly
differnet from each other."
## [1] "### T-test results of Sepal.Length and Petal.Length #########################
###########"
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 13.098, df = 211.54, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.771500 2.399166
## sample estimates:
## mean of x mean of y
##  5.843333  3.758000
##
## [1] "P value is smaller than (or equal to) 0.05 and thus we can reject the null hy
pothesis.  Consequently, the means of Sepal.Length and Petal.Length are significantly
differnet from each other."
## [1] "### T-test results of Sepal.Length and Petal.Width ##########################
##########"
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 50.536, df = 295.98, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.46315 4.82485
## sample estimates:
## mean of x mean of y
##  5.843333  1.199333
##
## [1] "P value is smaller than (or equal to) 0.05 and thus we can reject the null hy
pothesis.  Consequently, the means of Sepal.Length and Petal.Width are significantly
differnet from each other."
## [1] "### T-test results of Sepal.Width and Petal.Length ##########################
##########"
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -4.7194, df = 167.1, p-value = 4.975e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9937746 -0.4075587
## sample estimates:
## mean of x mean of y 
##  3.057333  3.758000 
## 
## [1] "P value is smaller than (or equal to) 0.05 and thus we can reject the null hy
pothesis.  Consequently, the means of Sepal.Width and Petal.Length are significantly 
differnet from each other."
## [1] "### T-test results of Sepal.Width and Petal.Width ###########################
#########"
## 
##  Welch Two Sample t-test
## 
## data:  x and y
## t = 25.916, df = 237.03, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.716763 1.999237
## sample estimates:
## mean of x mean of y 
##  3.057333  1.199333 
## 
## [1] "P value is smaller than (or equal to) 0.05 and thus we can reject the null hy
pothesis.  Consequently, the means of Sepal.Width and Petal.Width are significantly d
iffernet from each other."
## [1] "### T-test results of Petal.Length and Petal.Width ##########################
##########"
## 
##  Welch Two Sample t-test
## 
## data:  x and y
## t = 16.297, df = 202.69, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.249107 2.868227
## sample estimates:
## mean of x mean of y 
##  3.758000  1.199333 
## 
## [1] "P value is smaller than (or equal to) 0.05 and thus we can reject the null hy
pothesis.  Consequently, the means of Petal.Length and Petal.Width are significantly 
differnet from each other."
```