Lab1

Jonathan Stuart

2/13/2018

First, let's load the data set.

```
load("data/KaiserBabies.rda")
```

Next we can create variables for the population size and the sample size. The population size will be 1236 of all pregnancies that occurred between 1960 and 1967 among women in the Kaiser Foundation Health Plan in Oakland, CA.

```
population_size <- 1236
sample_size <- 10
```

Next we can take a simple random sample, without replacement, of size 10 from the population of 1236.

```
#taking a simple random sample
set.seed(7)
my_sample = sample(na.omit(infants$wt),10)
```

Now we can begin answering the questions asked in the lab assignment.

Question 1a

Use the sample average to estimate the average weight of the mothers.

```
#finding sample average
x_bar <- mean(my_sample)
```

Calculate the estimated standard error of these estimates.

```
#calculating s
s <- sqrt((sum((my_sample - x_bar) ^ 2)) / (sample_size - 1))

#calculating standard error
standard_error <- (s / sqrt(10)) *
  sqrt((population_size - sample_size) / (population_size - 1))
```

```
## Our estimated standard error is 4.894749.
```

Assuming normality holds, form a 95% confidence interval for the average of the population .

```
#constucting a confidence interval
lower_limit <- x_bar - 1.96 * standard_error
upper_limit <- x_bar + 1.96 * standard_error
```

```
## Our 95% confidence interval for average weight of mothers is [ 125.1063 , 144.2937 ]
```

Quesion 1b

Without using the `set.seed()` function, repeat this process 1000 times in order to create 1000 different confidence intervals.

```
#creating relevant variables
my_samples <- matrix(nrow = 1000, ncol = 10)
x_bars <- 0
standard_error_values <- 0
confidence_intervals <- matrix(nrow = 1000, ncol = 2)
confidence_intverval_accuracy <- matrix(nrow = 1000, ncol = 1)
ci_accuracy <- matrix(nrow = 1000, ncol = 1)
true_average <- mean(na.omit(infants$wt))

#for loop to generate confidence intervals
for (i in 1:1000) {
  my_samples[i, ] = sample(na.omit(infants$wt),10)
  x_bars[i] <- mean(my_samples[i, ])
  standard_error_values[i] <- sd(my_samples[i, 1:10]) / sqrt(10)
  confidence_intervals[i, 1] <- x_bars[i] - 1.96 * standard_error_values[i]
  confidence_intervals[i, 2] <- x_bars[i] + 1.96 * standard_error_values[i]
  ci_accuracy[i] <- (true_average <= confidence_intervals[i, 2] &
                       true_average >= confidence_intervals[i, 1])
}
```

How many of them do you expect to cover the true average?

Based on the definition of confidence intervals as the range that will contain the true value of the corresponding parameter with a specified degree of certainty, we would theoretically expect around 950 of them to cover the true average.

How many do?

```
## Of our 1000 sample averages, 113 values fell outside of their corresponding
## confidence intervals, while 887 values fell within their corresponding
## confidence intervals.
```

Since the sample size, 10, is quite small, it should not be surprising that the actual frequency of confidence intervals that cover the true average deviates from the theoretical expectation.

Question 1c

Calculate the SD of the sample averages.

```
#calculating the SD of the sample means
sd_means <- sd(x_bars)
sd_means
```

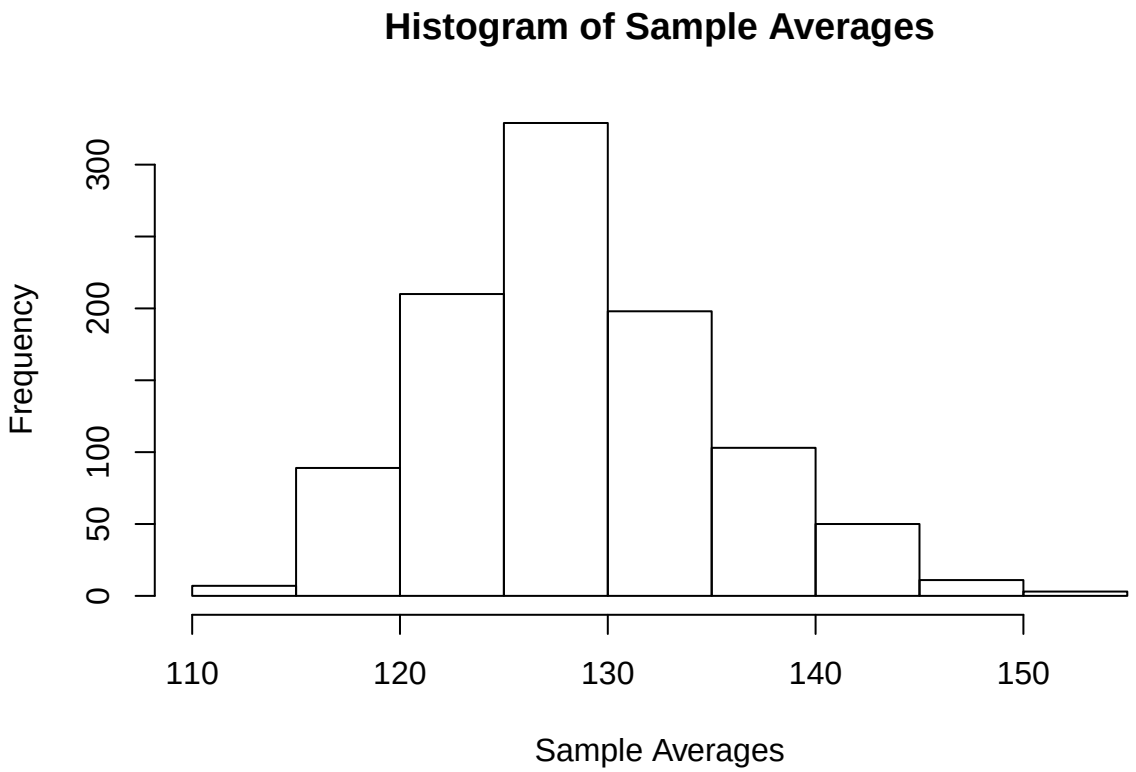
```
## [1] 6.740372
```

Is it close to the estimated standard error from Question 1a?

```
## The SD of our sample average is 6.740372, while the estimated standard error from
## Question 1a is 4.894749, a difference of 1.845623.
```

Make a histogram of the sample averages to see if it seems plausible that the probability histogram for the sample average follows the normal curve pretty closely.

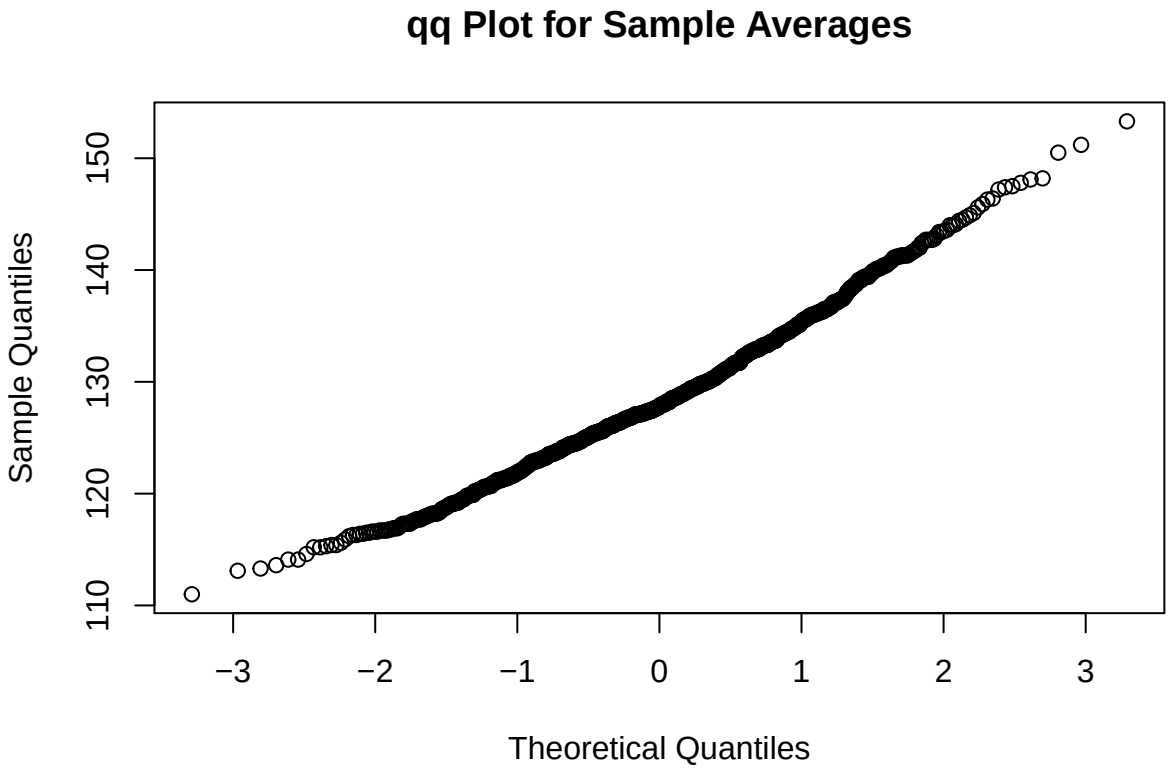
```
#histogram of sample averages
hist(x_bars, main = "Histogram of Sample Averages", xlab = "Sample Averages")
```



Yes, the histogram for the sample averages is approximately normal, and skewed slightly left with a slightly longer right tail.

Make a quantile-quantile plot to further investigate.

```
#qqplot of x_bars
qqnorm(x_bars, main = "qq Plot for Sample Averages")
```



Does it seem like the confidence interval is valid?

```
## It appears that slightly less that 95% of the mass of sample averages is contained
## within the interval [125.1063, 144.2937], perhaps 85-90%. This means that the
## confidence interval is slightly invalid.
```

Question 2a

Start with your original sample and use it to construct a bootstrap population.

```
#constructing the bootstrap population
vals = sort(unique(my_sample))
counts = table(my_sample)
# makes the bootstrap pop as rounded version of sample, not quite right
boot_pop <- rep(vals, round(counts * population_size / length(my_sample)))
length(boot_pop)
```

```
## [1] 1239
```

```
## boot_pop is now a vector of 1239, the number of members in the bootstrap population.
```

Using that bootstrap population, get 1000 simple random samples of size 10.

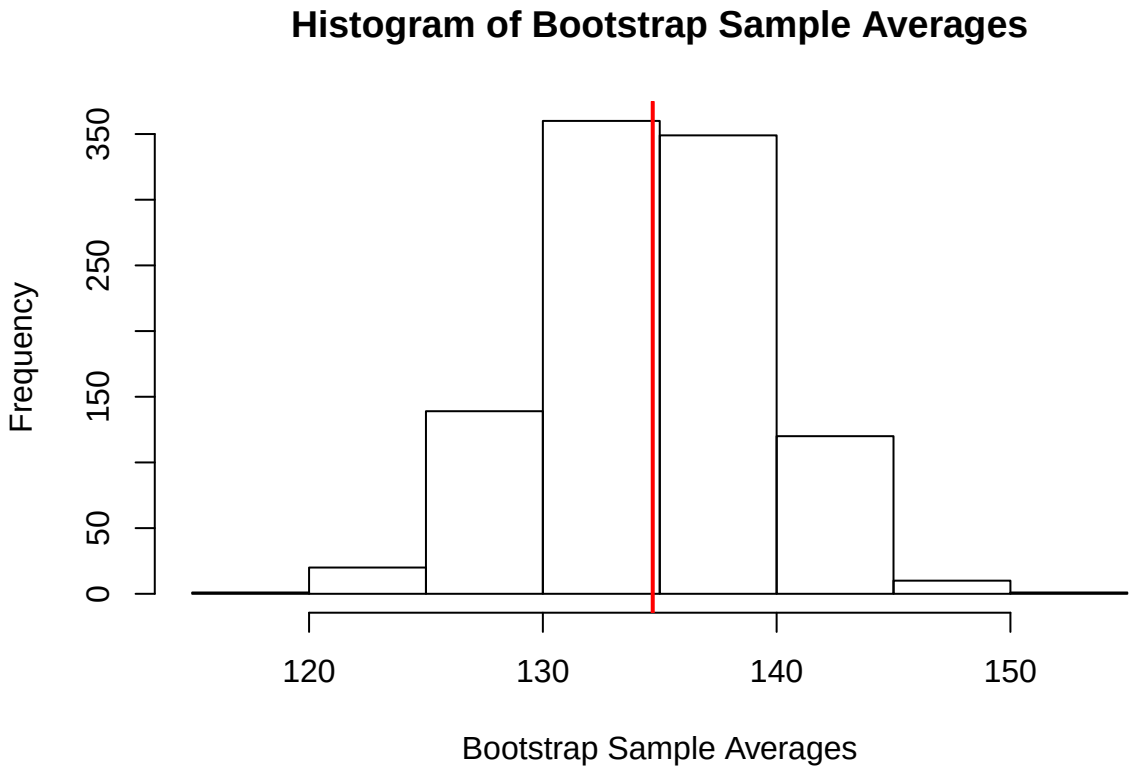
```
#sampling from the bootstrap population
boot_pop_sample <- replicate(1000, sample(boot_pop, length(my_sample), FALSE))
```

```
## boot_pop_sample is now a matrix of 10 rows (the number of members per sample),
## and 1000 columns (the total number of samples).
```

For each of the samples, calculate the sample average and make a histogram of these sample averages, putting a vertical line through the average of the bootstrap population.

```
#calculating the sample averages of the bootstrap samples
boot_x_bars <- 0
for (i in 1:1000) {
  boot_x_bars[i] <- mean(boot_pop_sample[ , i])
}

#histogram of sample averages with vertical line at bootstrap pop. average
hist(boot_x_bars, main = "Histogram of Bootstrap Sample Averages",
     xlab = "Bootstrap Sample Averages")
boot_pop_mean <- mean(boot_pop)
abline(v=boot_pop_mean,col="red", lwd = 2)
```



Calculate the SD of the sample averages.

```
#calculating the sd of the sample averages
standard_deviation_averages <- sd(boot_x_bars)
```

Is it close to the estimated standard error from Question 1a above?

```
## The SD of our sample averages is 4.71299, while the estimated standard error from
## Question 1a is 4.894749, a difference of 0.1817591.
```

So, yes! The values are very close.

Question 2b

Construct a 95% bootstrap confidence interval by taking the 2.5 percentile and the 97.5 percentile of the bootstrap sample averages.

```
#getting the quantlies of the bootstrap sample averages
bootstrap_quantile <- quantile(boot_x_bars, probs = seq(0, 1, .025))

#taking a look at the `bootstrap_quantile` vector
head(bootstrap_quantile)
```

```
##       0%     2.5%       5%     7.5%      10%    12.5%
## 119.6000 125.2975 127.2000 128.0925 128.6000 129.2000
```

```
## The 95% confidence interval for the bootstrap population is [ 125.2975 ,  143.5025 ]
```

How does it compare to the confidence interval you got in Question 1a?

```
## The 95% confidence interval associated with our bootstrap population is
## [125.2975, 143.5025], while the 95% confidence interval associated with our estimated
## standard error in Question 1a was [125.1063, 144.2937]. The two intervals are nearly
## equivalent.
```

As we can see, the confidence intervals are very, very close.

Now, we will replicate the same procedure for a sample size of 100 instead of a sample size of 10.

```
population_size <- 1236
sample_size <- 100
```

Next we can take a simple random sample, without replacement, of size 100 from the population of 1236.

```
#taking a simple random sample
set.seed(7)
my_sample = sample(na.omit(infants$wt),100)
```

Now we can begin answering the questions asked in the lab assignment.

Question 1a

Use the sample average to estimate the average weight of the mothers.

```
#finding sample average
x_bar <- mean(my_sample)
```

Calculate the estimated standard error of these estimates.

```
#calculating s
s <- sqrt((sum((my_sample - x_bar) ^ 2)) / (sample_size - 1))

#calculating standard error
standard_error <- (s / sqrt(100)) *
  sqrt((population_size - sample_size) / (population_size - 1))
```

```
## Our estimated standard error is 1.974992.
```

Assuming normality holds, form a 95% confidence interval for the average of the population .

```
#constucting a confidence interval
lower_limit <- x_bar - 1.96 * standard_error
upper_limit <- x_bar + 1.96 * standard_error
```

```
## Our 95% confidence interval for average weight of mothers is [ 125.909 , 133.651 ]
```

Quesion 1b

Without using the `set.seed()` function, repeat this process 1000 times in order to create 1000 different confidence intervals.

```
#creating relevant variables
my_samples <- matrix(nrow = 1000, ncol = 100)
x_bars <- 0
standard_error_values <- 0
confidence_intervals <- matrix(nrow = 1000, ncol = 2)
confidence_intverval_accuracy <- matrix(nrow = 1000, ncol = 1)
ci_accuracy <- matrix(nrow = 1000, ncol = 1)
true_average <- mean(na.omit(infants$wt))

#for loop to generate confidence intervals
for (i in 1:1000) {
  my_samples[i, ] = sample(na.omit(infants$wt),100)
  x_bars[i] <- mean(my_samples[i, ])
  standard_error_values[i] <- sd(my_samples[i, 1:100]) / sqrt(100)
  confidence_intervals[i, 1] <- x_bars[i] - 1.96 * standard_error_values[i]
  confidence_intervals[i, 2] <- x_bars[i] + 1.96 * standard_error_values[i]
  ci_accuracy[i] <- (true_average <= confidence_intervals[i, 2] &
                       true_average >= confidence_intervals[i, 1])
}
```

How many of them do you expect to cover the true average?

Based on the definition of confidence intervals as the range that will contain the true value of the corresponding parameter with a specified degree of certainty, we would theoretically expect around 950 of them to cover the true average.

How many do?

```
## Of our 1000 sample averages, 39 values fell outside of their corresponding
## confidence intervals, while 961 values fell within their corresponding
## confidence intervals.
```

Since the sample size, 100, is much larger than 10, it makes sense that the actual frequency of confidence intervals that cover the true average is closer to the theoretical expectation.

Question 1c

Calculate the SD of the sample averages.

```r
#calculating the SD of the sample means
sd_means <- sd(x_bars)
sd_means
```

```
## [1] 2.000529
```

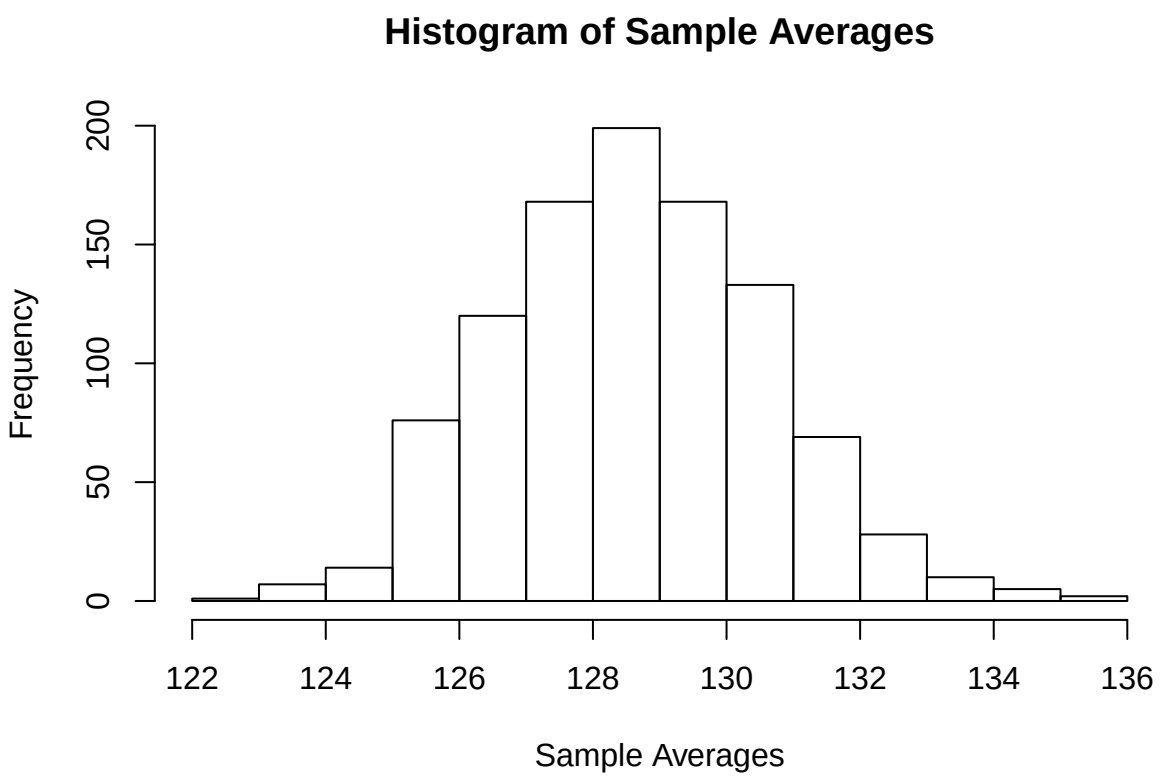
Is it close to the estimated standard error from Question 1a?

```
## The SD of our sample average is 2.000529, while the estimated standard error from
## Question 1a is 1.974992, a difference of 0.0255372.
```

Yes, the values are very close.

Make a histogram of the sample averages to see if it seems plausible that the probability histogram for the sample average follows the normal curve pretty closely.

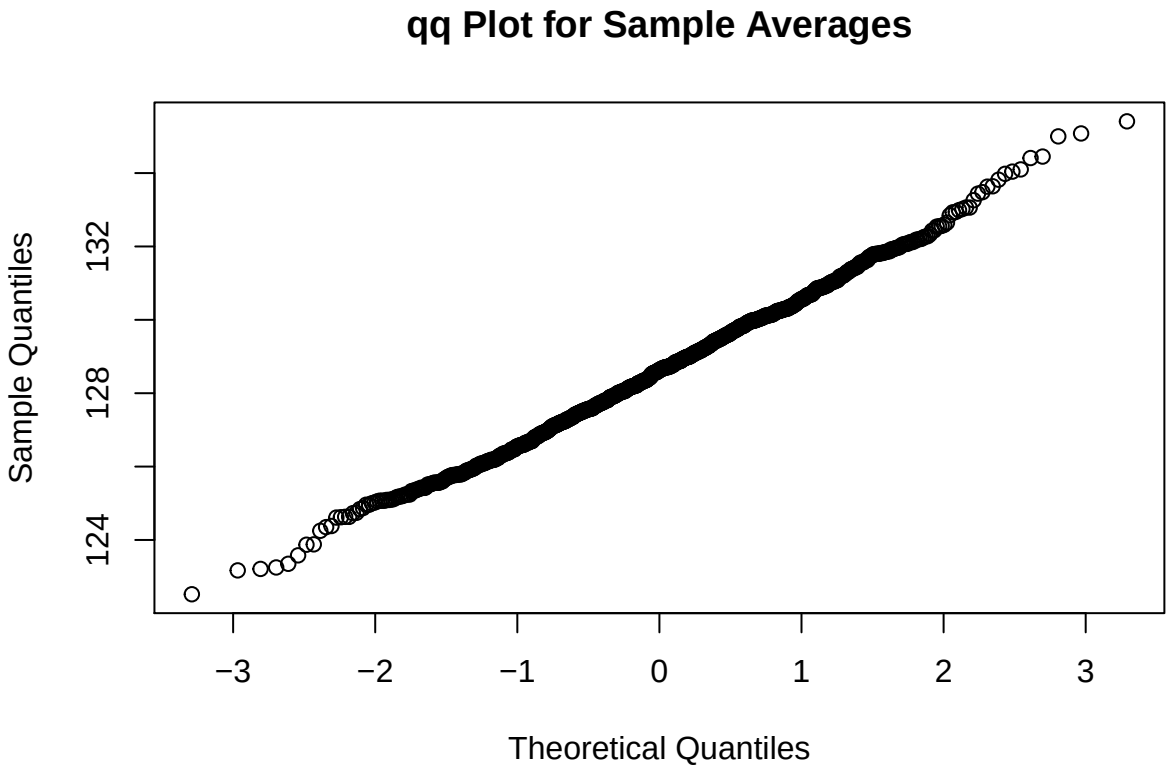
```r
#histogram of sample averages
hist(x_bars, main = "Histogram of Sample Averages", xlab = "Sample Averages")
```



Yes, the histogram for the sample averages is very nearly normal, if not outright normal. The histogram very closely follows the normal curve.

Make a quantile-quantile plot to further investigate.

```r
#qqplot of x_bars
qqnorm(x_bars, main = "qq Plot for Sample Averages")
```



Does it seem like the confidence interval is valid?

```
## It appears that very nearly 95% of the mass of sample averages is contained
## within the interval [125.909, 133.651]. This means that the
## confidence interval is extremely valid.
```

Question 2a

Start with your original sample and use it to construct a bootstrap population.

```
#constructing the bootstrap population
vals = sort(unique(my_sample))
counts = table(my_sample)
# makes the bootstrap pop as rounded version of sample, not quite right
boot_pop <- rep(vals, round(counts * population_size / length(my_sample)))
length(boot_pop)
```

```
## [1] 1232
```

```
## boot_pop is now a vector of 1232, the number of members in the bootstrap population.
```

Using that bootstrap population, get 1000 simple random samples of size 100.

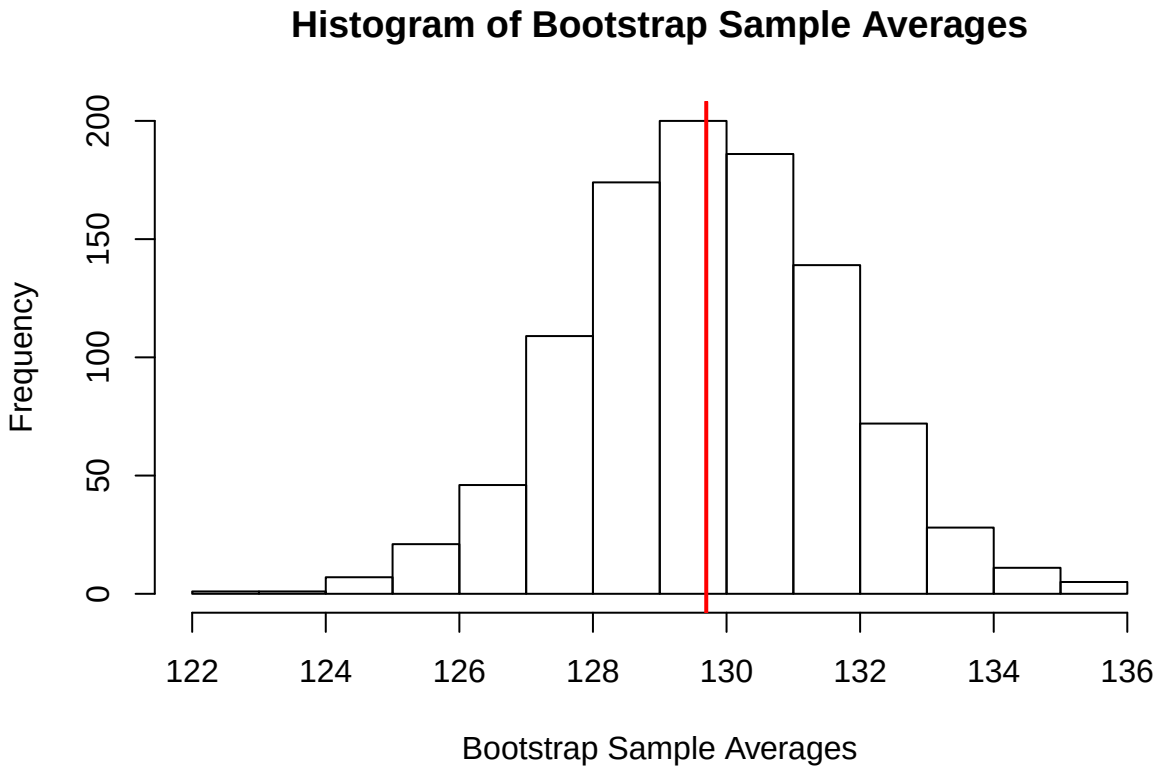
```
#sampling from the bootstrap population
boot_pop_sample <- replicate(1000, sample(boot_pop, length(my_sample), FALSE))
```

```
## boot_pop_sample is now a matrix of 100 rows (the number of members per sample),
## and 1000 columns (the total number of samples).
```

For each of the samples, calculate the sample average and make a histogram of these sample averages, putting a vertical line through the average of the bootstrap population.

```
#calculating the sample averages of the bootstrap samples
boot_x_bars <- 0
for (i in 1:1000) {
  boot_x_bars[i] <- mean(boot_pop_sample[ , i])
}

#histogram of sample averages with vertical line at bootstrap pop. average
hist(boot_x_bars, main = "Histogram of Bootstrap Sample Averages",
     xlab = "Bootstrap Sample Averages")
boot_pop_mean <- mean(boot_pop)
abline(v=boot_pop_mean,col="red", lwd = 2)
```



Calculate the SD of the sample averages.

```
#calculating the sd of the sample averages
standard_deviation_averages <- sd(boot_x_bars)
```

Is it close to the estimated standard error from Question 1a above?

```
## The SD of our sample averages is 1.951793, while the estimated standard error from
## Question 1a is 1.974992, a difference of 0.02319852.
```

So, yes! The values are very close.

Question 2b

Construct a 95% bootstrap confidence interval by taking the 2.5 percentile and the 97.5 percentile of the bootstrap sample averages.

```
#getting the quantlies of the bootstrap sample averages
bootstrap_quantile <- quantile(boot_x_bars, probs = seq(0, 1, .025))

#taking a look at the `bootstrap_quantile` vector
head(bootstrap_quantile)
```

```
##       0%     2.5%       5%     7.5%      10%    12.5%
## 122.6400 125.8898 126.5895 126.9892 127.2690 127.4975
```

```
## The 95% confidence interval for the bootstrap population is [ 125.8898 ,  133.6702 ]
```

How does it compare to the confidence interval you got in Question 1a?

```
## The 95% confidence interval associated with our bootstrap population is
## [125.8898, 133.6702], while the 95% confidence interval associated with our estimated
## standard error in Question 1a was [125.909,133.651]. The two intervals are nearly
## equivalent.
```

As we can see, the confidence intervals are very, very close.