Math 490 HW #15

Maxwell Levin

March 21, 2018

Question 1.

A chi-square distribution with a low degree of freedom is strongly right-skewed. In R, we can use the command rchisq(15, 1) to generate 15 random numbers (a SRS of size 15) from a chi-square distribution with 1 degree of freedom. This exercise explores the effects of sample size and non-normality on the bootstrap sampling distribution.

a. Use R to draw a SRS of size 15 from a chi-square population with 1 degree of freedom. Make a histogram of this data and bootstrap the sample mean $\overline{X}$ using $B = 1024$ resamples from your original sample. Make a histogram of the bootstrap sampling distribution and report the bootstrap standard error. Briefly compare the two histograms.

We can run the following code in R to draw a SRS from a chi-square distribution:

```
chiSq15 = rchisq(15, 1)
step = (min(chiSq15) + max(chiSq15))/20    # Just to make the plot look nicer
hist(chiSq15, breaks=seq(min(chiSq15), max(chiSq15) + step, step))
```

Histogram of chiSq15
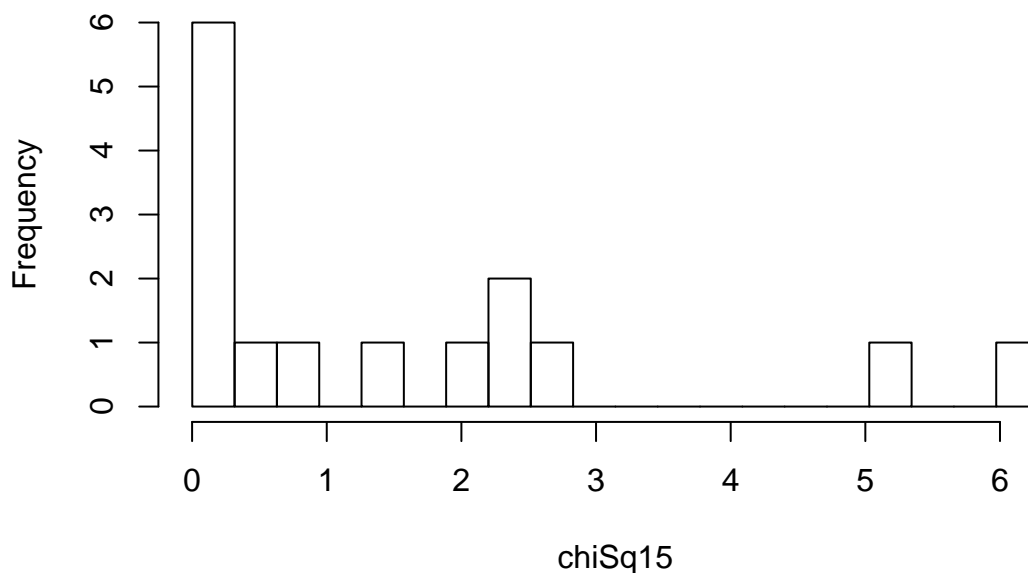
We now bootstrap the sample mean and make a histogram by running the following code in R:

```
bootMean = function(data, rep) {
  len = length(data)
  replicate(rep, mean(sample(data, len, replace=TRUE)))
```
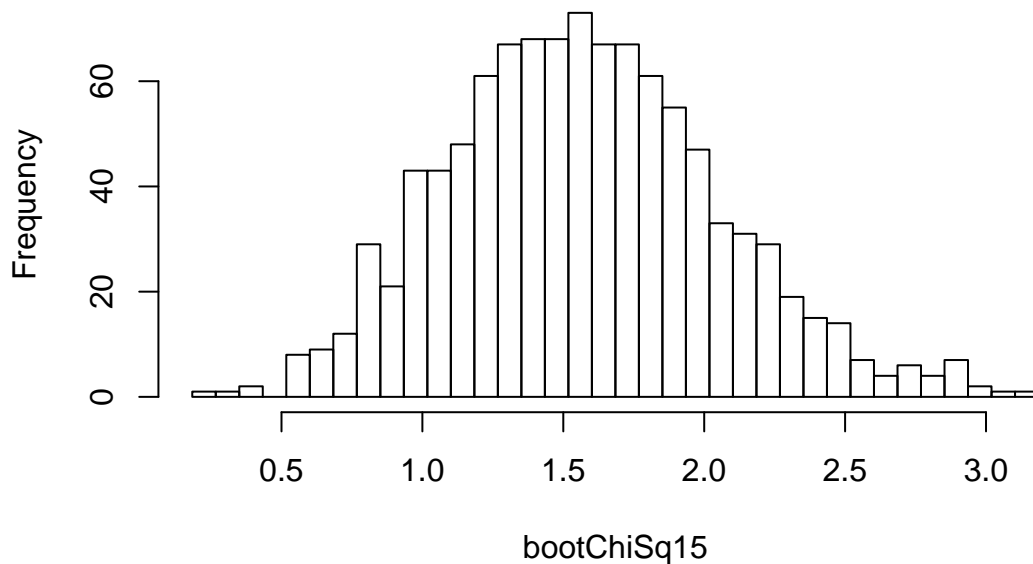
```
}

bootChiSq15 = bootMean(chiSq15, 1024)
step = (min(bootChiSq15) + max(bootChiSq15))/40
hist(bootChiSq15, breaks=seq(min(bootChiSq15), max(bootChiSq15) + step, step))
```

Histogram of bootChiSq15

The standard deviation of our bootstrap distribution is

```
[1] 0.4783274
```

Our two histograms do not look very similar in shape, but they are both centered around 1.5. The bootstrapped distribution appears much more normal than our sample distribution, and our sample distribution appears like it might be skewed to the right, although it is hard to tell with so few data points.

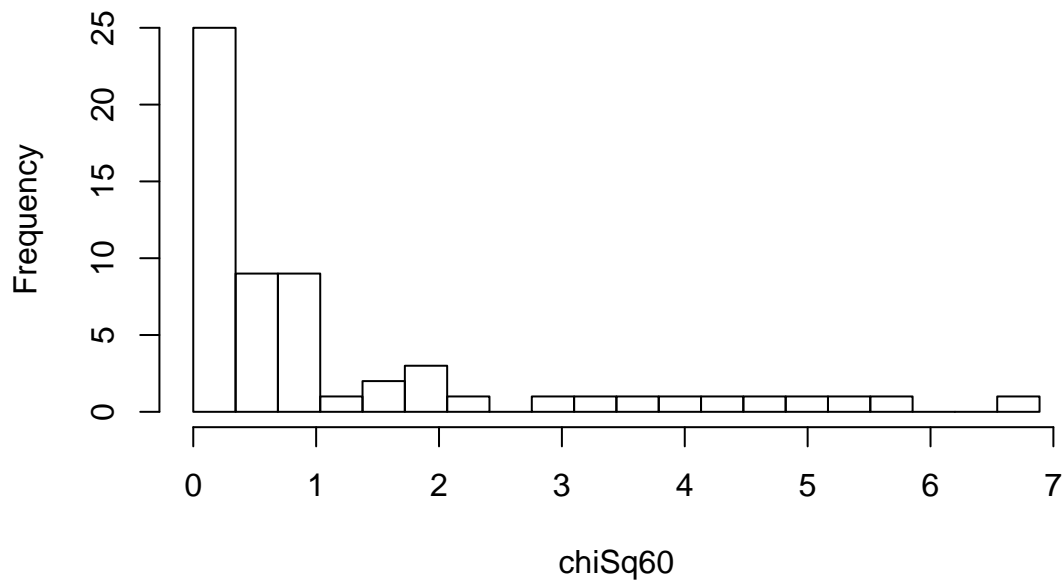**b. Repeat the process in (a) for a SRS of size 60.**

We can run the following code in R to draw a SRS from a chi-square distribution:

```
chiSq60 = rchisq(60, 1)
step = (min(chiSq60) + max(chiSq60))/20
hist(chiSq60, breaks=seq(min(chiSq60), max(chiSq60) + step, step))
```
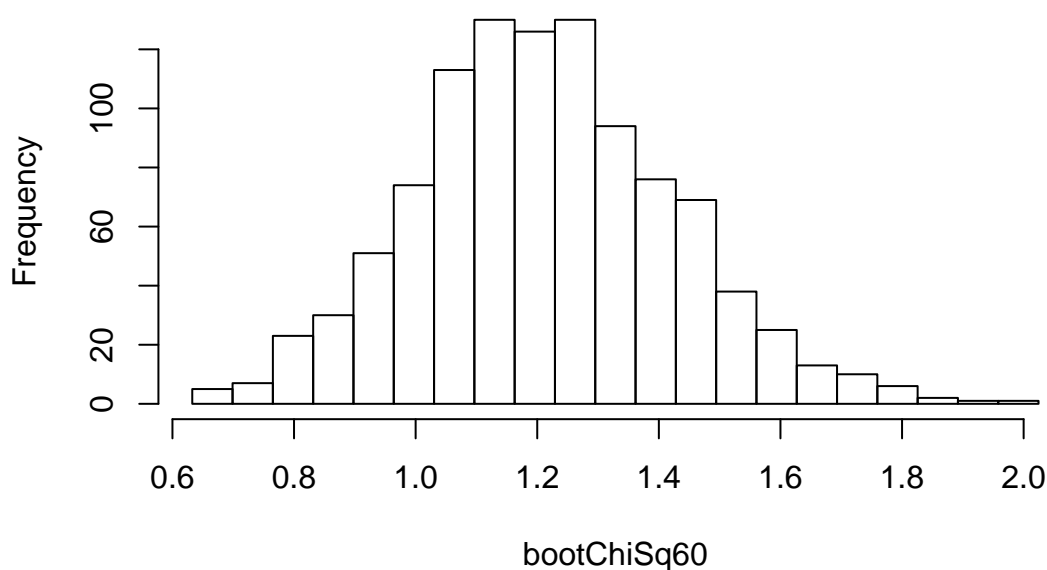
Histogram of chiSq60

We now bootstrap the sample mean by running the following code in R:

```r
bootMean = function(data, rep) {
  len = length(data)
  replicate(rep, mean(sample(data, len, replace=TRUE)))
}


bootChiSq60 = bootMean(chiSq60, 1024)
step = (min(bootChiSq60) + max(bootChiSq60))/40
hist(bootChiSq60, breaks=seq(min(bootChiSq60), max(bootChiSq60) + step, step))
```

Histogram of bootChiSq60

The standard error of our bootstrapped mean distribution is

```
[1] 0.2130841
```

Although our two histograms both appear skewed to the right, the sample histogram appears much more skewed than our bootstrapped distribution. Our bootstrapped distribution appears relatively normal, and has a much smaller spread than our sample histogram. The two histograms could feasibly both centered around 0.9.
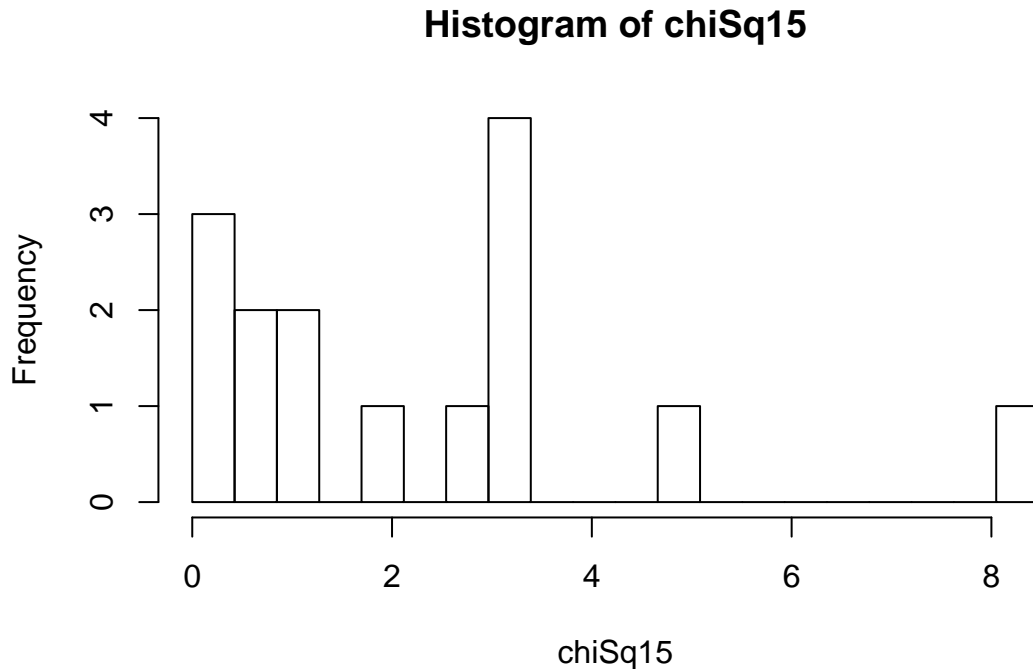
c. Write a brief paragraph comparing the two bootstrap sampling distributions in (a) and (b). Be sure to discuss the effects of increasing the sample size.

The bootstrap sampling distribution in (a) looks pretty similar to the bootstrap sampling distribution in (b) in terms of shape, but the bootstrap sampling distribution in (a) has a much larger spread than in (b) (and thus a higher standard error) in addition to being centered at a slightly higher value than the distribution in (b). I think the reason for this is that our sample data that we bootstrap from in (a) is of lower quality than our sample data in part (b). I believe this is in part due to the difference in sample sizes, as our sample size increases our data becomes more and more like the actual distribution we are studying. Thus a bootstrap distribution from a large sample is more likely to be better representative of our actual datathan a bootstrap distribution from a small sample.

d. Use R to draw a SRS of size 15 with a mean of 1 and a variance of 2. Make a histogram of this and then bootstrap the sample mean $\overline{X}$ using $B = 1024$ resamples from your original sample. Make a histogram of the bootstrap sampling distribution and report the bootstrap standard error. Briefly compare the two histograms.

We can run the following code in R to draw a SRS from a chi-square distribution:

```
chiSq15 = rchisq(15, 1, sqrt(2))
step = (min(chiSq15) + max(chiSq15))/20    # Just to make the plot look nicer
hist(chiSq15, breaks=seq(min(chiSq15), max(chiSq15) + step, step))
```
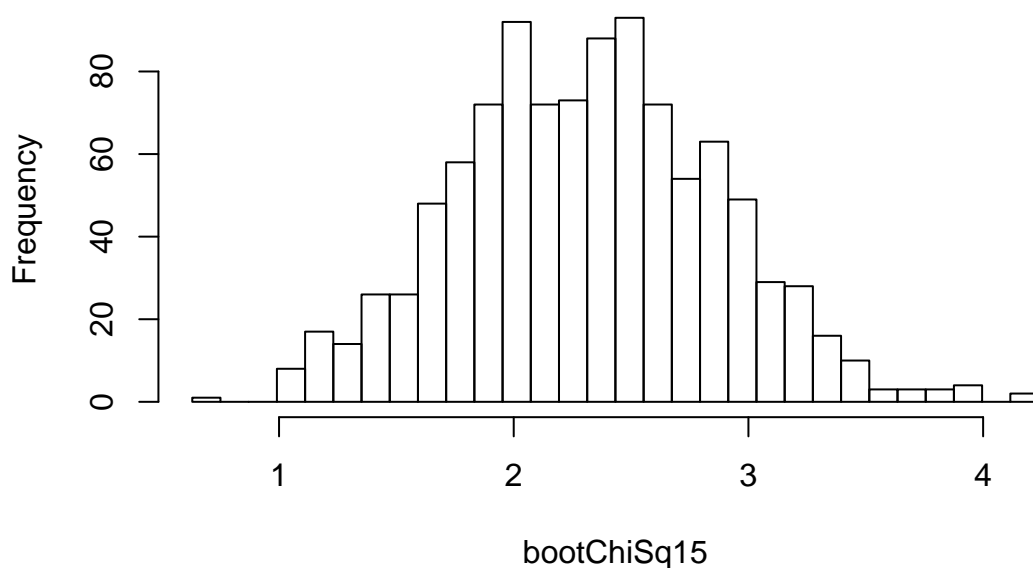
Histogram of chiSq15

We now bootstrap the sample mean by running the following code in R:

```
bootMean = function(data, rep) {
  len = length(data)
  replicate(rep, mean(sample(data, len, replace=TRUE)))
}


bootChiSq15 = bootMean(chiSq15, 1024)
step = (min(bootChiSq15) + max(bootChiSq15))/40
hist(bootChiSq15, breaks=seq(min(bootChiSq15), max(bootChiSq15) + step, step))
```

Histogram of bootChiSq15

The bootstrap standard error is

```
[1] 0.5537391
```

As with the other bootstrap sample distributions we've looked at so far in this assignment, the bootstrap sample distribution here looks much more normally distributed than our sample distribution, and they are similar in terms of center. In addition, it appears that the spread of our sample distribution is much larger than the spread of our bootstrap distribution.

Question 2.

The distribution of the 60 IQ test scores is roughly normal and the sample size is large enough that we expect a normal sampling distribution. We will compare confidence intervals for the population mean IQ, μ based on this sample.

a. Use the formula $s/\sqrt{n}$ to find the standard error of the mean. Give the 95% t-confidence interval based on this standard error.

We use the plug-in method to substitute our sample standard deviation for s and then divide by the square root of 60 to get:

```
[1] 1.910792
```

Thus our sample standard error of the mean is about 1.91. Our 95% confidence interval based on this statistic is given by

$$Y \pm t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.$$

Since we know from the last homework assignment that the mean IQ is about 114.98, this becomes

$$114.98 \pm t_{n-1, \frac{\alpha}{2}} (1.910792),$$

$$[111.16, 118.81].$$

This is our 95% confidence interval.

b. Bootstrap the mean of the IQ scores. Make a histogram and normal quantile plot of the bootstrap distribution. Does the bootstrap distribution appear normal? What is the bootstrap standard error? Give the bootstrap 95% t-confidence interval.

We can run the following code in R to bootstrap the mean of the IQ scores:

```
bootMean = function(data, rep) {
  len = length(data)
  replicate(rep, mean(sample(data, len, replace=TRUE)))
}

simMeanIQ = bootMean(IQ, 1000)
hist(simMeanIQ, breaks=seq(min(simMeanIQ), max(simMeanIQ) + 0.5, 0.5), probability=TRUE)
```
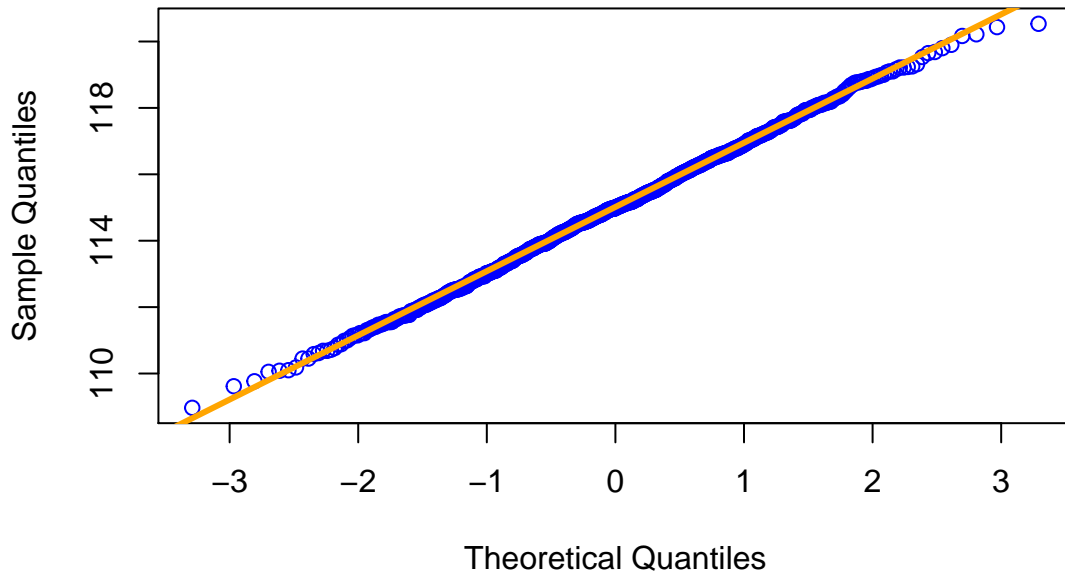
Histogram of simMeanIQ
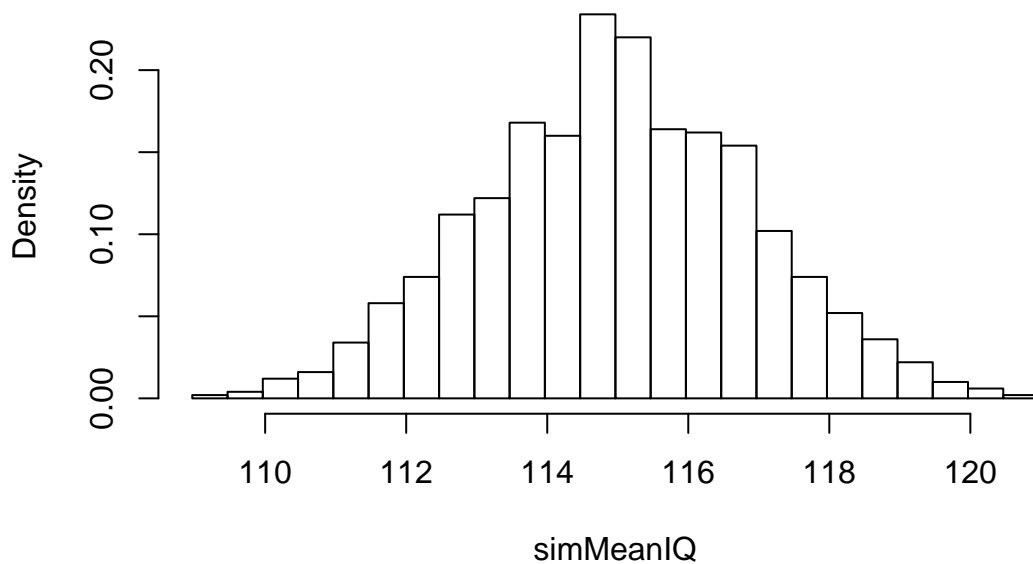
This looks fairly normal, but lets check with a normal quantile plot:

```
qqnorm(simMeanIQ, col="blue")
qqline(simMeanIQ, col="orange", lwd=3)
```

Normal Q–Q Plot

This does indeed appear to be normally distributed.

The bootstrap standard error is

```
[1] 1.935566
```

We follow a similar process here to get a 95% confidence interval:

$$Y \pm t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}},$$

Where we use the mean of our bootstrap distribution for Y and the standard error for $\frac{s}{\sqrt{n}}$. In R this becomes:

```
lower = mean(simMeanIQ) - abs(qt(.975, length(simMeanIQ) - 1) * sd(simMeanIQ))
upper = mean(simMeanIQ) + abs(qt(.975, length(simMeanIQ) - 1) * sd(simMeanIQ))
lower
```

```
[1] 111.1906
```
```
upper
```

```
[1] 118.7871
```

Thus we see that the 95% confidence interval for our bootstrap means is about $[111.1, 118.6]$.

c. Give the 95% confidence percentile. Make a graphical comparison by drawing a vertical line at the original sample mean $\bar{x}$ and displaying the other five intervals horizontally, one above the other. How well do your confidence intervals agree? Was bootstrapping needed to find a reasonable confidence interval, or was the formula-based confidence interval good enough?

We can compute the bootstrap 95% confidence percentile by running the following code in R:
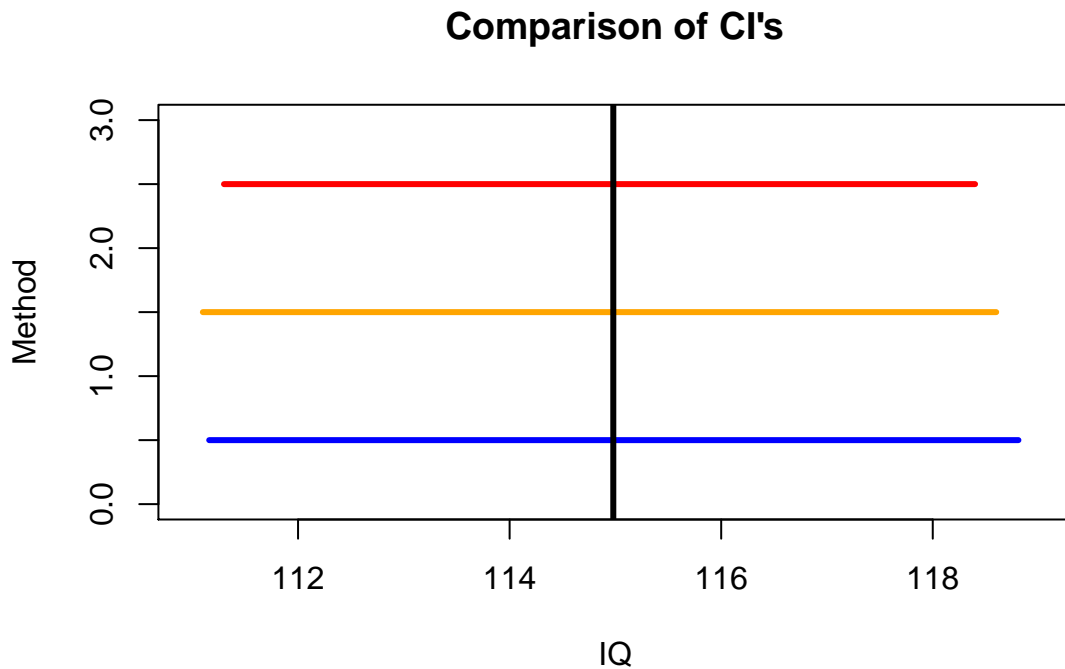
```
quantile(simMeanIQ, 0.025)
```

```
    2.5%
111.2167
```

```
quantile(simMeanIQ, 0.975)
```

```
   97.5%
118.8179
```

Thus the 95% bootstrap percentile confidence interval is approximately [111.3, 118.4].

Comparison of CI's

Here the blue horizontal line represents the 95% sample mean confidence interval, the orange line is the 95% bootstrap mean confidence interval, the red line is the 95% bootstrap percentile confidence interval, and the vertical black line represents the actual sample mean. As we can see, all of these intervals are centered at, or very close to, the sample mean. Their spreads are all very similar, and they start and end in just about the same places (there is a little bit of variablilty in the two bootstrap intervals due to random sampling). Since the two bootstrap confidence intervals are both very similar to the confidence interval from the original sample, I think that bootstrapping was probably not needed to find a reasonable confidence interval.