


## lab6andrews

Christopher Andrews

11/26/2018

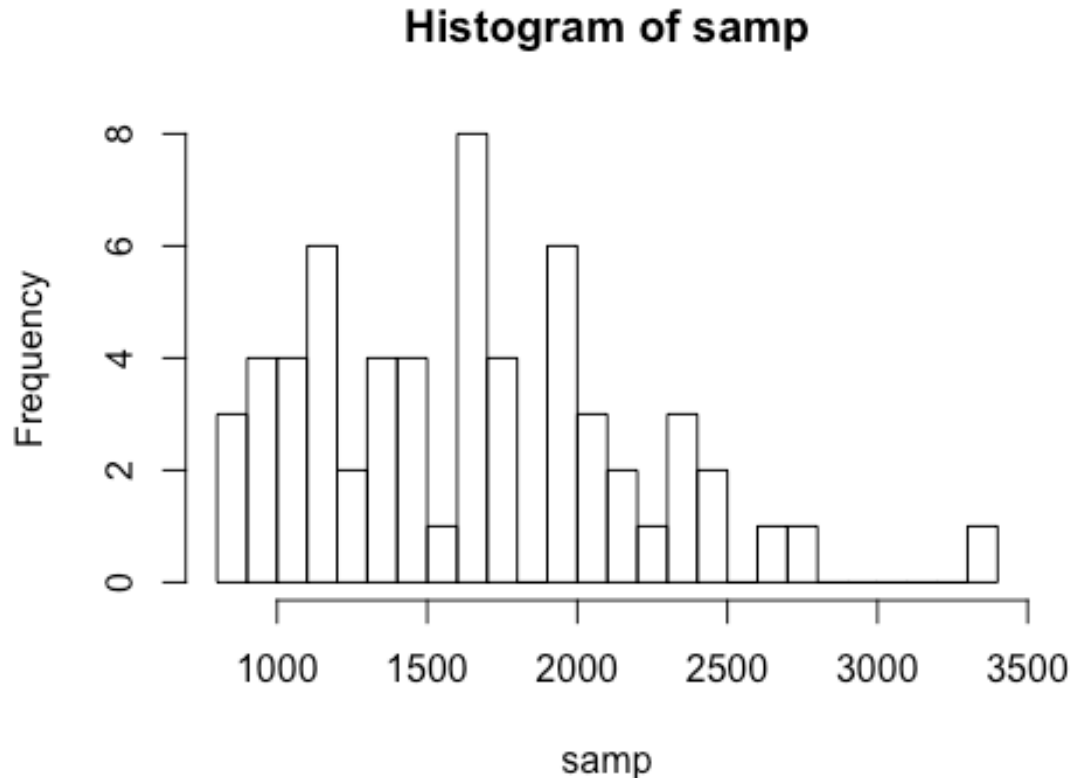
```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile =  
"ames.RData")  
load("ames.RData")
```

```
population <- ames$Gr.Liv.Area  
samp <- sample(population, 60)
```

 **Exercise 1:** Plot a histogram of your sample of living areas. Then, describe the shape, center, and spread of your histogram. What would you say is the “typical” living area within your sample? Explain.

The center is around 1607... The shape of the main part of the histogram is slightly right skewed... There looks to be at least one outlier around 3400... I would say the typical living area based on the sample is 1632.. The spread of this histogram is 545, and just by 'eye-balling' it, there seems to be a high outlier.

```
hist(samp, breaks = 25)
```



```
summary(samp)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.     Max.
##      810   1168   1607   1632   1975   3390
```

```
sd(samp)
```

```
## [1] 544.7877
```

**Exercise 2:** Would you expect another student's sample distribution to be identical to yours? Would you expect it to be similar? Why or why not?

No, I would not expect another student's sample distribution to be identical. I would expect it to be entirely different. But, there is definitely a small chance that the sampling distribution would be exactly the same. I would expect the sampling another student's sample distribution to be somewhat similar.

```
sample_mean <- mean(samp)
```

```
se <- sd(samp)/sqrt(60)
```

```
lower <- sample_mean - 2 * se
```

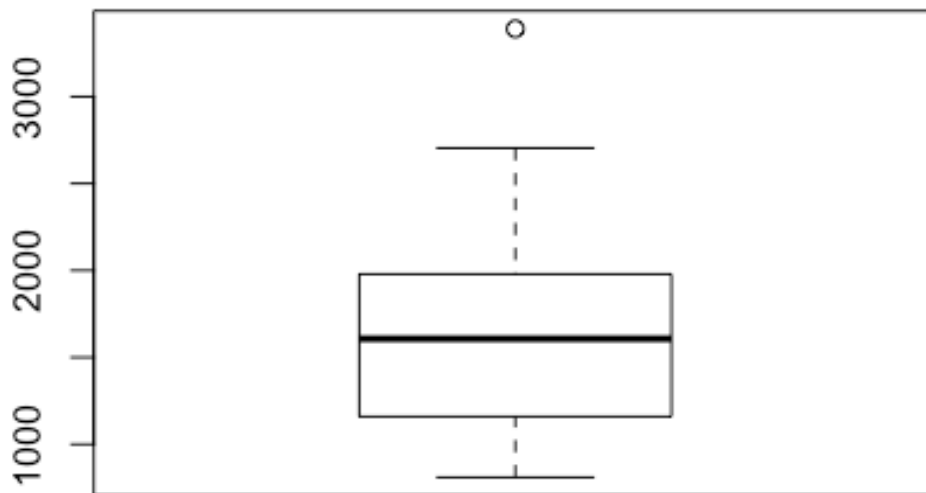
```
upper <- sample_mean + 2 * se
```

```
c(lower, upper)
```

```
## [1] 1491.553 1772.880
```

**Exercise 3:** For a one-sample t confidence interval to be valid, the sampling distribution of the sample mean must be normally distributed. Check this assumption using the indirect methods demonstrated during class. (Note: If any outliers are present in your sample, you will need to include the relevant calculations to classify the outlier(s) as being either mild or extreme. Extreme outliers prevent us from applying the Central Limit Theorem.)

```
boxplot(samp)
```

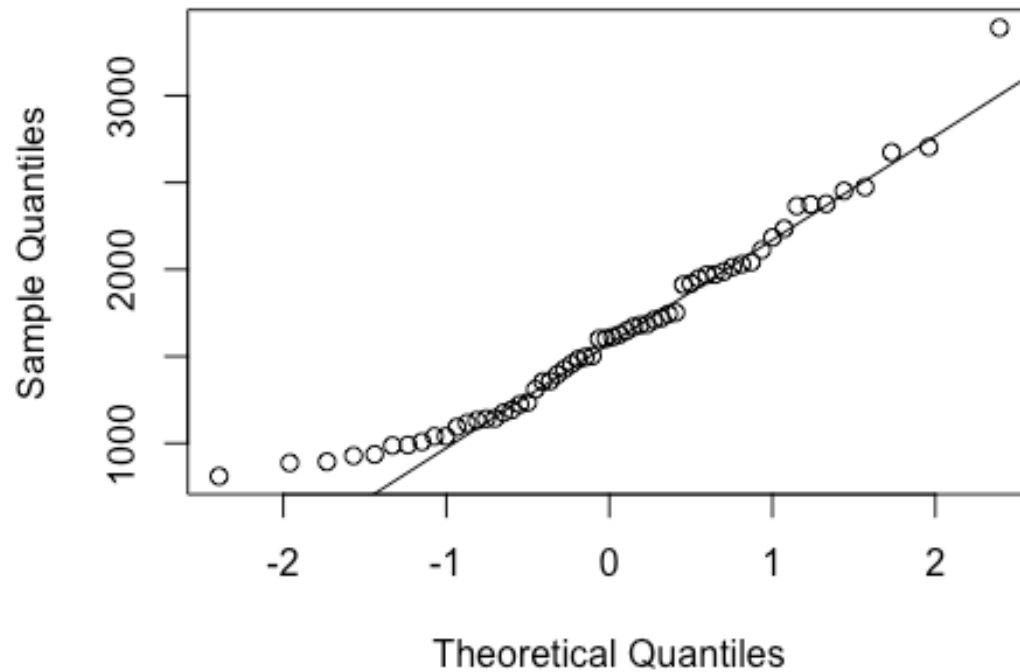


```
shapiro.test(samp)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  samp  
## W = 0.95337, p-value = 0.0225
```

```
qqnorm(samp, main = "QQ plot")  
qqline(samp)
```

## QQ plot



```
summary(samp)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      810   1168   1607   1632   1975   3390
```

```
IQR(samp)
```

```
## [1] 807.25
```

```
extremeHigh <- 1975 + (IQR(samp) * 3)
```

```
mildHigh <- 1975 + (IQR(samp) * 1.5)
```

```
extremeLow <- 1168 - (IQR(samp) * 3)
```

```
mildLow <- 1168 - (IQR(samp) * 1.5)
```

```
extremeHigh
```

```
## [1] 4396.75
```

```
mildHigh
```

```
## [1] 3185.875
```

```
mildLow
```

```
## [1] -42.875
extremeLow
## [1] -1253.75
```

For the histogram, it is right skewed by a little bit. The sample has no low outliers, based on the minimum is 810 and the fence for mild low outliers is -42.875. There is however a mild high outlier. Our max, 3390 is greater than the mild high fence for outliers (3185.875). Just by looking at the box plot as well as the histogram, you can see that there is at least one outlier on the high side. Now to the Shapiro-Wilk test, our p value is less than the p value for normality (.25), while our p-value is 0.0225.

➡ **Exercise 4:** Report your 95% confidence interval in the form . Then, carefully interpret your confidence interval in context.

$1491.553 < \mu < 1772.880$  We are 95% confident that the true mean of living area of houses in this data set is between 1491.553 and 1772.880.

➡ **Exercise 5:** What does the phrase “95% confident” mean? In other words, give an interpretation of the confidence level.

It means that we are 95% confident that in between the lower and upper limit, lies the population mean. So, when confidence decreases (less than 95%) the interval gets larger. Vice Versa, when the confidence interval grows (greater than 95%) the interval gets smaller. In this example, the true population mean does actually lie between our lower and upper limit.

$1491.553 < 1499.69 < 1772.880$

➡ **Exercise 6:** Did your confidence interval capture the true mean living area of houses in Ames? Explain.

Yes my confidence interval did capture the true mean living area of houses in Ames. (  $1491.553 < 1499.69 < 1772.880$  ) Based on the above questions answer and this data demonstrated on the line above, the confidence interval we produced does capture the true mean.

➡ **Exercise 7:** Each student in your class section should have gotten a slightly different confidence interval. What proportion of those intervals would you expect to successfully capture the true population mean? Why? Write your confidence interval on the board. When everybody has done so, write down the confidence intervals created by all of the students in your class section and calculate the proportion of these intervals that successfully captured the true population mean. How does this proportion compare to the expected proportion? Why might it be different? Explain.

```
43 * 0.95
## [1] 40.85
```

Out of the data we have, there were 43 submitted confidence intervals. 95 percent of 43 is 40.85, so we can round up to 41. Based on this calculation I would expect 41 submitted confidence intervals to capture the true population mean. It actually turns out that 42 out of 43 submitted confidence intervals capture the true population mean.

```
samp_mean <- rep(NA, 50)
samp_sd <- rep(NA, 50)
n <- 60

for(i in 1:50){
  samp <- sample(population, n) # obtain a sample of size n = 60 from the
population
  samp_mean[i] <- mean(samp) # save sample mean in ith element of samp_mean
  samp_sd[i] <- sd(samp) # save sample sd in ith element of samp_sd
}

lower <- samp_mean - 2 * samp_sd/sqrt(n)
upper <- samp_mean + 2 * samp_sd/sqrt(n)

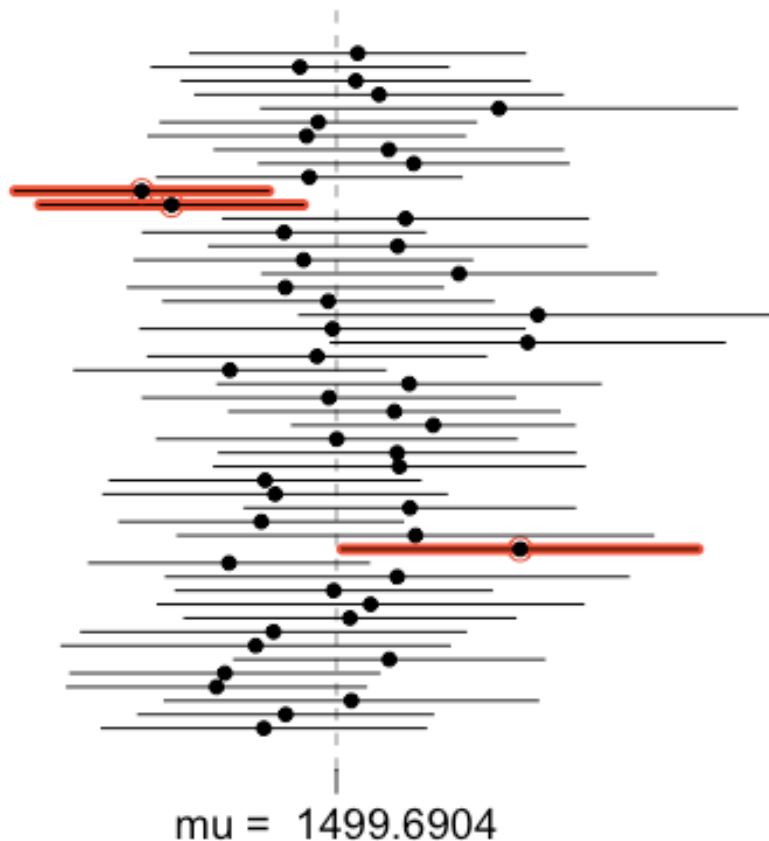
c(lower[1], upper[1])

## [1] 1323.253 1567.014
```

## Homework Assignment

**1. Using the following function (which was downloaded with the data set), plot all fifty of your 95% confidence intervals:**

```
plot_ci(lower, upper, mean(population))
```



What proportion of your confidence intervals include the true population mean? Is this proportion exactly equal to the confidence level? Why might it differ?

47 of the 50 obtain the true population mean. This is just about equal to the confidence level. 95 percent of 50 is 47.5, so if we rounded down, then it is equal. But it is very close to be exactly 95%. This may differ because it is a confidence interval of a sample of the true population mean, not the actual true population mean.

2. **What is the appropriate critical t value for a 98% confidence level with 59 df? Include R calculations for finding this critical t. (It could be helpful to also find the critical t using the invT command on your graphing calculator. Confirm that you get the same result using both methods to ensure that you used the correct R command.)**

```
qt(.99,59)
```

```
## [1] 2.391229
```

3. Construct fifty 98% confidence intervals. You do not need to obtain new samples; simply calculate new intervals based on the sample means and standard deviations you have already collected; you only need to change the critical t used in the calculations (it was 2 for a 95% confidence level and 59 df). Using the `plot_ci` function, plot all fifty intervals and calculate the proportion of intervals that include the true population mean. How does this percentage compare to the confidence level?

```
newMean <- rep(NA, 50)
newSD <- rep(NA, 50)
num <- 60

for(i in 1:50){
  newSamp <- sample(population, num) # obtain a sample of size n = 60 from the
population
  newMean[i] <- mean(newSamp) # save sample mean in ith element of samp_mean
  newSD[i] <- sd(newSamp) # save sample sd in ith element of samp_sd
}

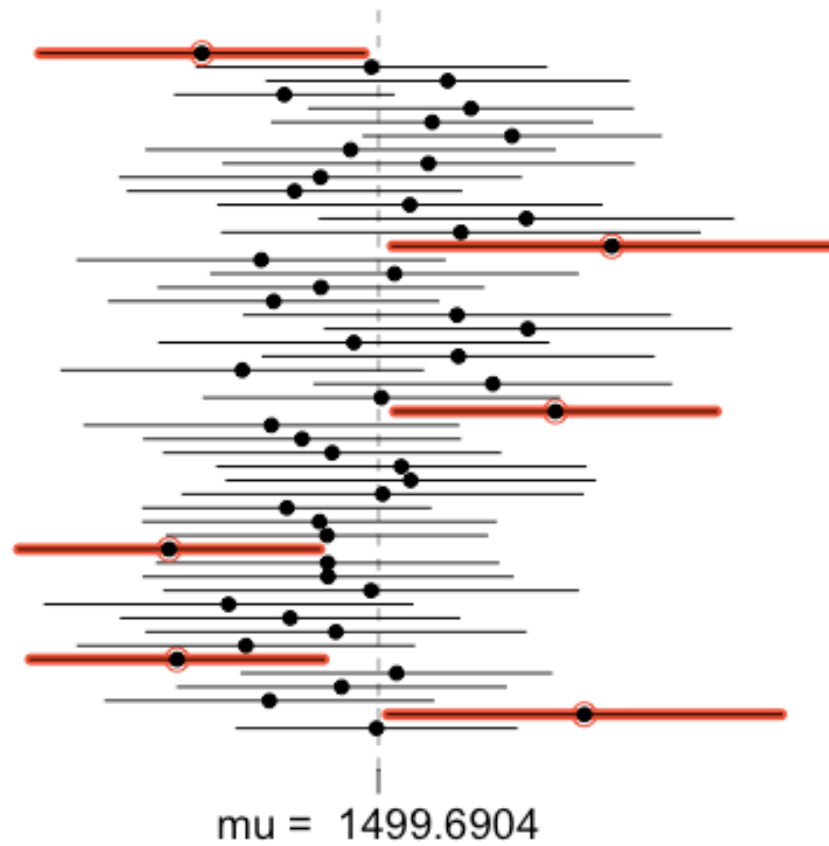
newLower <- newMean - 2 * newSD/sqrt(num)
newUpper <- newMean + 2 * newSD/sqrt(num)

c(newLower[1], newUpper[1])

## [1] 1399.109 1597.424

plot_ci(newLower, newUpper, mean(population))
```





49 out of 50 intervals contain the true population mean. This is exactly 98%.