Lab 6: The Confidence Level

Quentin Terry

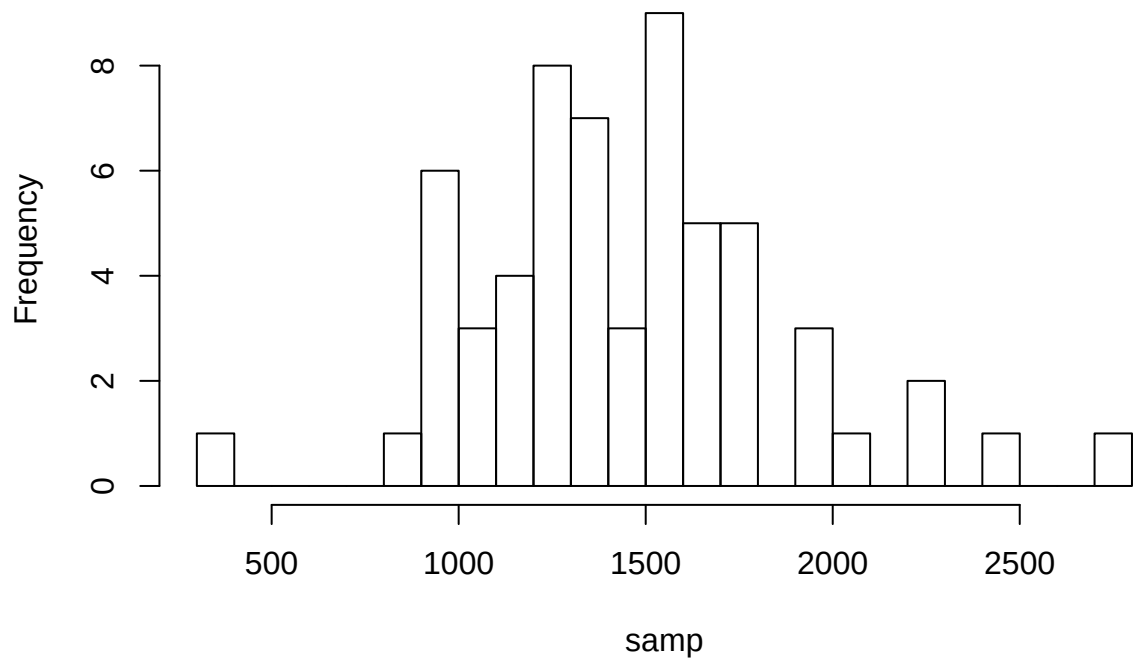
November 19, 2018

```
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.

```
hist(samp, main = "Histogram of Sample Living Areas", breaks = 20)
```




```
summary(samp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     334    1206    1412    1450    1661    2730
```

```r
sd(samp)
```

```
## [1] 412.8894
```

-It can be seen that this histogram is slightly left skewed. The center of the histogram is 1450 and the spread is 413 with a high outlier. The typical living area within my sample is 1450 due to the fact that that is the sample mean of my sample.

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?

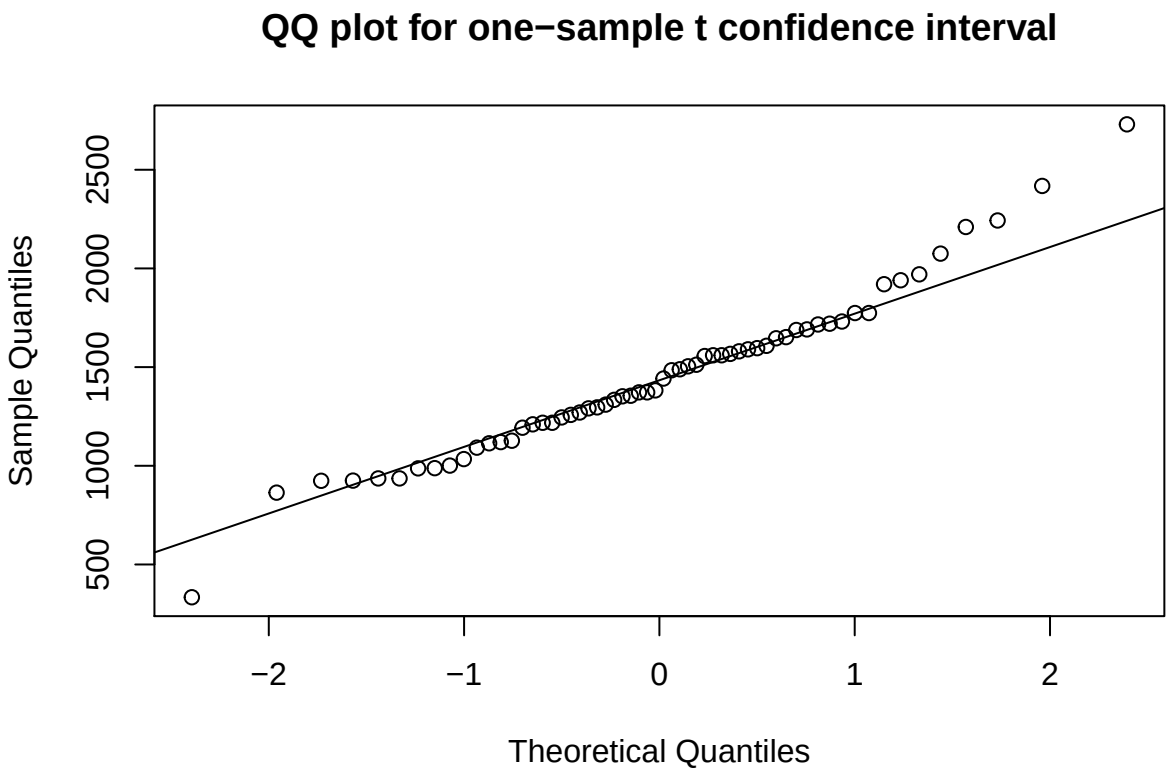
I dont think another student's sample distribution would be identical to mine, however I do expect it to be similar. This is due to the fact that no student could have the same seed as me, however the values are similar.

```r
sample_mean <- mean(samp)

se <- sd(samp)/sqrt(60)
lower <- sample_mean - 2 * se
upper <- sample_mean + 2 * se
c(lower, upper)
```

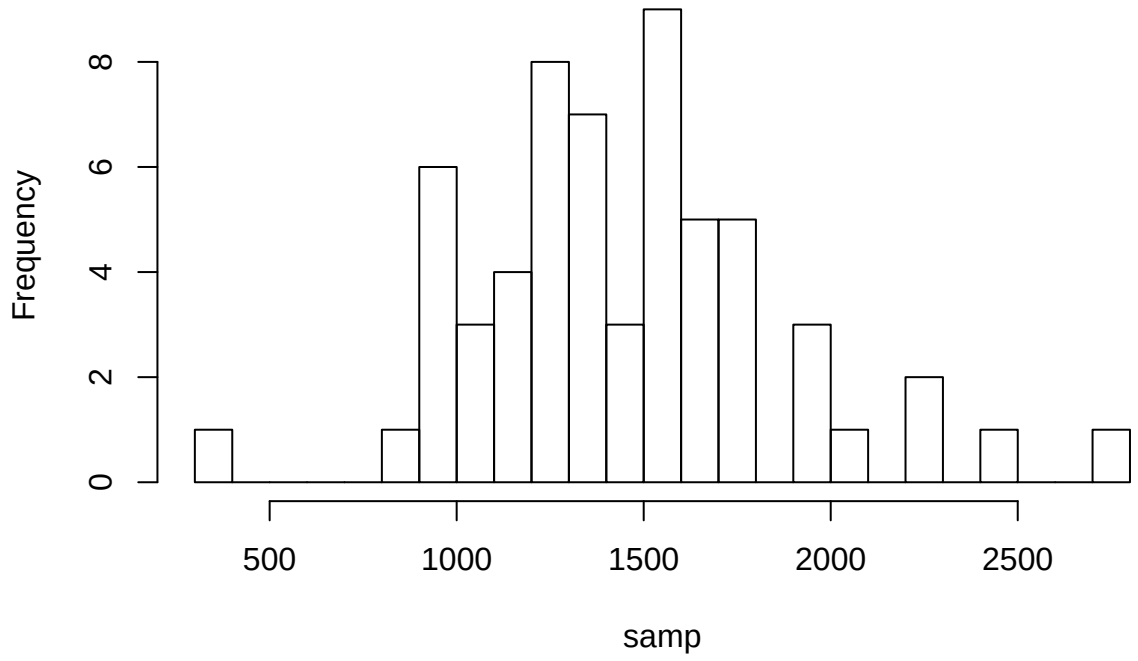
```
## [1] 1343.459 1556.674
```

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.)

```r
qqnorm(samp, main = "QQ plot for one-sample t confidence interval")
qqline(samp)
```




```r
hist(samp, breaks = 20,main = "Histogram of one-sample t confidence interval")
```




```
summary(samp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     334    1206    1412    1450    1661    2730
```

```
IQR(samp)
```

```
## [1] 455.25
```

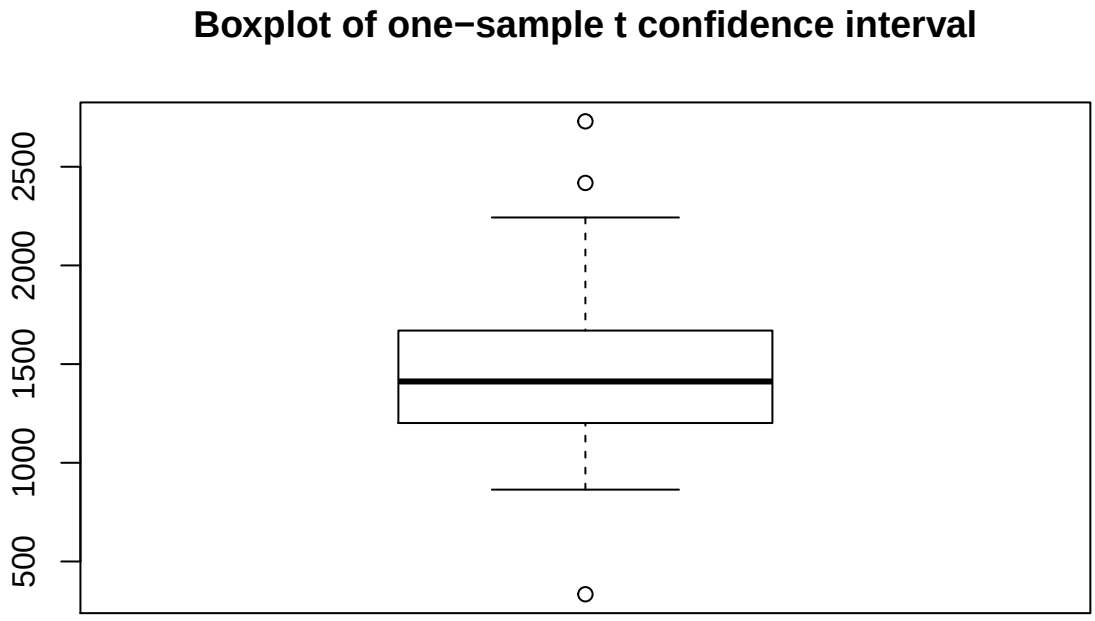
```
#Extreme high = Q3 + 3 * IQR

#Extreme low = Q1 - 3 * IQR

1661 + 3 * 455
```

```
## [1] 3026
```

```
1206 - 3 * 455
```

```
## [1] -159
```

```
boxplot(samp, main ="Boxplot of one-sample t confidence interval")
```




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

```
##
##  Shapiro-Wilk normality test
##
## data:  samp
## W = 0.97068, p-value = 0.1575
```

- The histogram for the sample is right-skewed, the box plot has no extreme outliers as $2730 < 3026$ and $-159 < 334$ and is not symetric. The qq plot is also right skewed. Also, in doing a shapiro wilk test we can see that the p value of the sample is less than the p value for normality, $1.575 < 0.25$. With this information it is reasonable to assume that under the central limit theorem, the sampling distribution of every possible sample mean is normal even though the the sample is not normal.

Exercise 4: Report your 95% confidence interval in the form x ??? E < μ < x + E. Then, carefully interpret your confidence interval in context.

$1343.459 < \mu < 1556.674$

We are 95% confident that the true mean living area of houses in Ames lies between 1343.459 and 1556.674.

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

```
mean(population)
```

```
## [1] 1499.69
```

The phrase "95% confident" means that we belive that the true population mean is most likely in between the confidence interval we gathered. As we can see, the true population mean is between the confidence intervals we derived with the sample at $1343.459 < 1499.69 < 1556.674$.

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

Yes, As we can see, the true population mean is between the confidence intervals we derived with the sample at $1343.459 < 1499.69 < 1556.674$. Therefore our confidence interval captured the true mean living area of houses in ames.

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

I would expect 41 out of 43 intervals to capture the true population mean. This is because each interval is a .95% confidence interval. Meaning that its 95% confident that its between the interval. The proportion of the class's intervals is 42 out of 43. This propotion is very close to the expected proportion. It may be different due to the fact that each student is doing an interval on a sample and the results may vary.

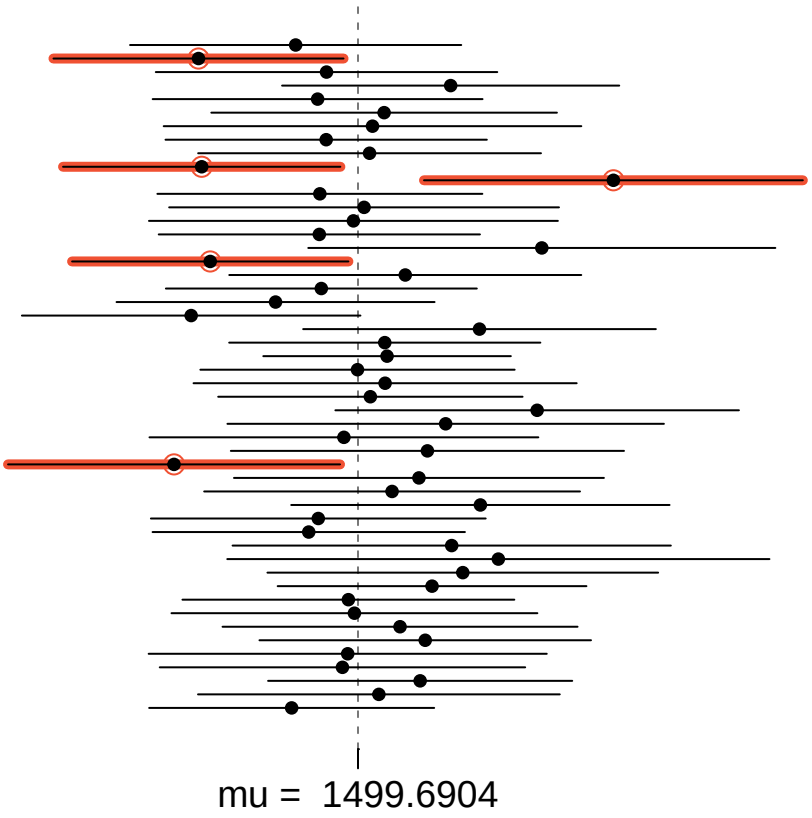
1. Using the following function (which was downloaded with the data set), plot all fifty of your 95% confidence intervals: plot_ci(lower, upper, mean(population)) What proportion of your confidence intervals include the true population mean? Is this proportion exactly equal to the confidence level? Why might it differ?

```
set.seed(001)

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] 1343.459 1556.674
```

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



- 45 out of 50 confidence intervals include the true population mean. This proportion is not exactly equal to the confidence level. This may differ because it is a 95% confident interval of a sample of the 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(0.99,59)
```

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

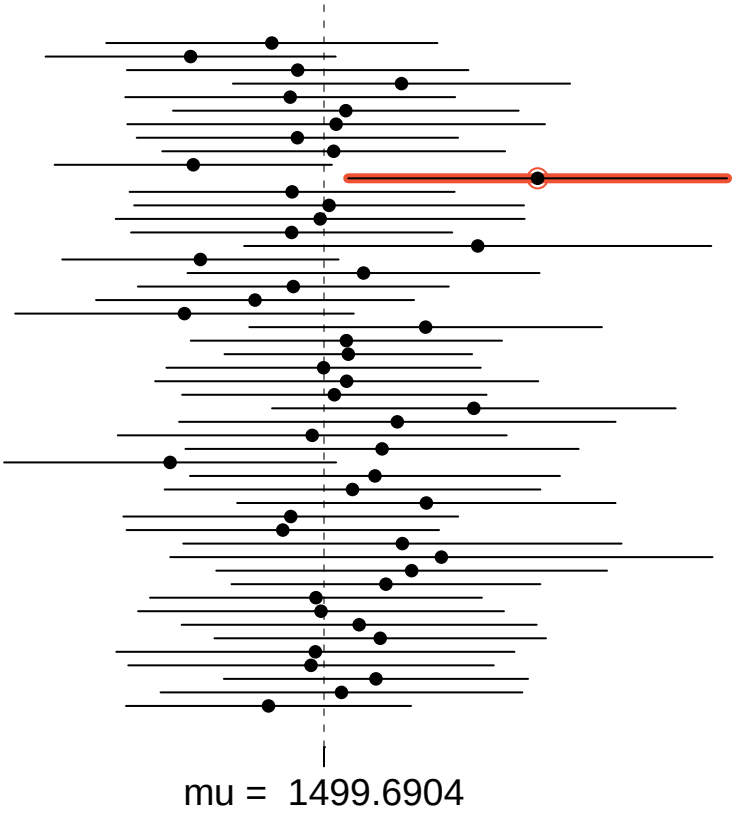
invT(0.99,59) -invT = 2.391229

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?

```
set.seed(001)

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 - 2.391229 * samp_sd/sqrt(n)
upper <- samp_mean + 2.391229 * samp_sd/sqrt(n)

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

```
## [1] 1322.605 1577.528
```

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



- The proportion of intervals that include the true population mean is 49 out of 50. This percentage the exact same as the confidence interval of 98%.