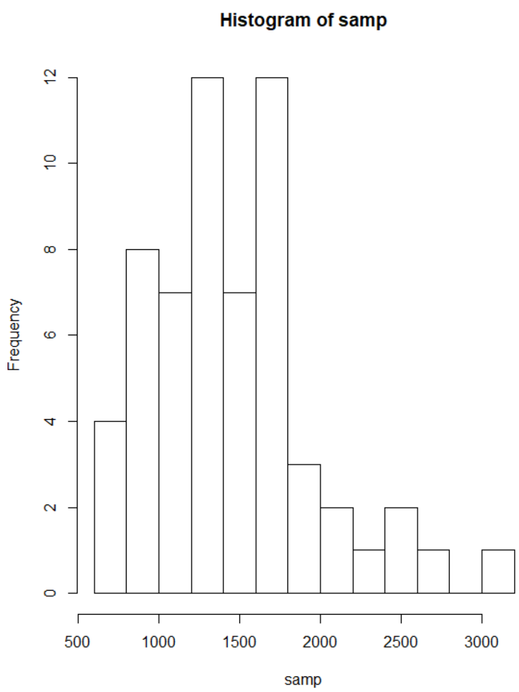
Name: Samuel L. Peoples

R Lab 5: Confidence Intervals

Exercise 1: Describe the distribution of your sample. What would you say is the "typical" size within your sample? Also state precisely what you interpreted "typical" to mean.




```
> hist(samp, breaks = 15)
> qqnorm(samp)
> qqline(samp)
> summary(samp)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
612     1042     1397    1447    1702    3082
```

This sample has a skewed-right distribution with a mean of 1447 and IQR of 660. The typical size within this sample is 1447, which is the average across those sampled.

Exercise 2: Would you expect another student's distribution to be identical to yours? Would you expect it to be similar? Why or why not?
We can expect other distributions to be similar but not identical. Because we took a random sample from the population, similar samples will have similar distributions.

Exercise 3: For the confidence interval to be valid, the sample mean must be normally distributed and have standard error √s/n. What conditions must be met for this to be true?
Samples must be independent and random, the sample size must be greater than 30, and normally distributed.

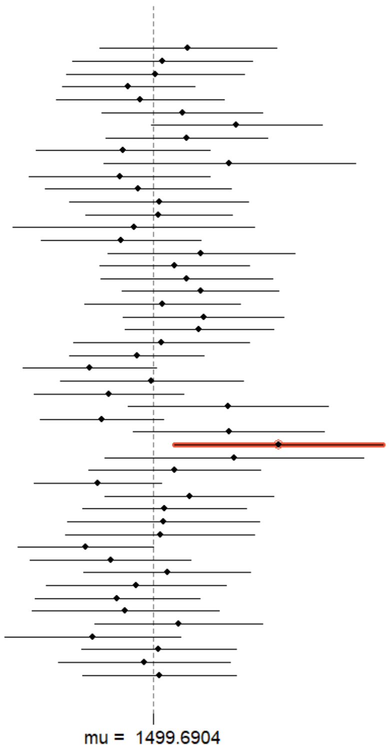
Exercise 4: What does "95% confidence" mean? If you're not sure, see Section 4.2.2.
A confidence of .95 estimates that 95% of the time, estimates will be within two standard deviations from the parameter. In this instance, the population parameter α=.05.

Exercise 5: Does your confidence interval capture the true average size of houses in Ames? If you are working on this lab in a classroom, does your neighbor's interval capture this value?

```
> sample_mean <- mean(samp)
> se <- sd(samp) / sqrt(60)
> lower <- sample_mean - 1.96 * se
> upper <- sample_mean + 1.96 * se
> c(lower, upper)
                                               1315.008 1578.359
> mean(population)
                                                         1499.69
```

The true mean of the population is within the 95% confidence interval.

Exercise 6: Each student in your class should have gotten a slightly different confidence interval. What proportion of those intervals would you expect to capture the true population mean? Why? If you are working in this lab in a classroom, collect data on the intervals created by other students in the class and calculate the proportion of intervals that capture the true population mean.



I would argue that 95% of the samples would capture the true mean.

On Your Own:

1. Using the following function (which was downloaded with the data set), plot all intervals. What proportion of your confidence intervals include the true population mean? Is this proportion exactly equal to the confidence level? If not, explain why.

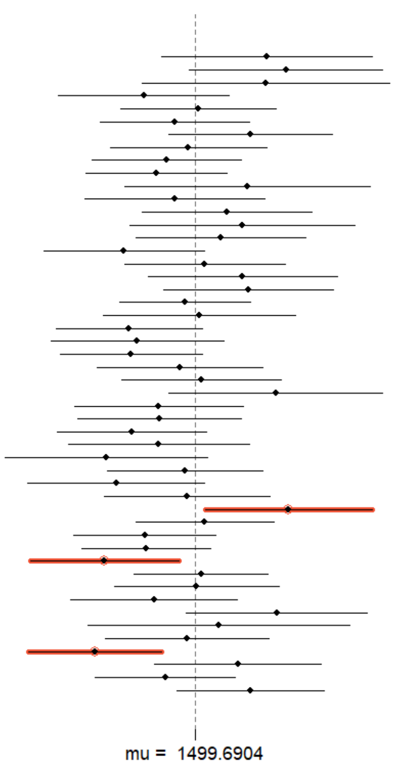
```
plot_ci(lower_vector, upper_vector, mean(population))
```

Out of fifty samples, there was only one sample which did not include the true population mean. This is a greater confidence interval, being 98%, and is different because the likelihood of the mean being in the interval is already quite high, and will level out at 95% over more intervals.

2. Pick a confidence level of your choosing, provided it is not 95%. What is the appropriate critical value?

```
> qnorm(.95)
```

For $\alpha = .1$, the confidence level is $1 - \frac{\alpha}{2} = .95$, and the critical value is 1.644.



3. Calculate 50 confidence intervals at the confidence level you chose in the previous question. You do not need to obtain new samples, simply calculate new intervals based on the sample means and standard deviations you have already collected. Using the plot_ci function, plot all intervals and calculate the proportion of intervals that include the true population mean. How does this percentage compare to the confidence level selected for the intervals?

```
> samp_mean <- rep(NA, 50)
> samp_sd <- rep(NA, 50)
> n <- 60
> for(i in 1:50){
+     samp <- sample(population, n)
+     samp_mean[i] <- mean(samp)
+     samp_sd[i] <- sd(samp)
+ }
> lower <- samp_mean - 1.645 * samp_sd / sqrt(n)
> upper <- samp_mean + 1.645 * samp_sd / sqrt(n)
> c(lower[1],upper[1])
                                    1476.512 1662.754
> plot_ci(lower, upper, mean(population))
```

In this interval, three out of the fifty samples did not contain the mean, which is roughly 94%. This is logical because we reduced our confidence from 5% to 10%.