Foundations for statistical inference - Confidence intervals

Sampling from Ames, Iowa

If you have access to data on an entire population, say the size of every house in Ames, Iowa, it's straight forward to answer questions like, "How big is the typical house in Ames?" and "How much variation is there in sizes of houses?". If you have access to only a sample of the population, as is often the case, the task becomes more complicated. What is your best guess for the typical size if you only know the sizes of several dozen houses? This sort of situation requires that you use your sample to make inference on what your population looks like.

The data

In the previous lab, "Sampling Distributions", we looked at the population data of houses from Ames, Iowa. Let's start by loading that data set.

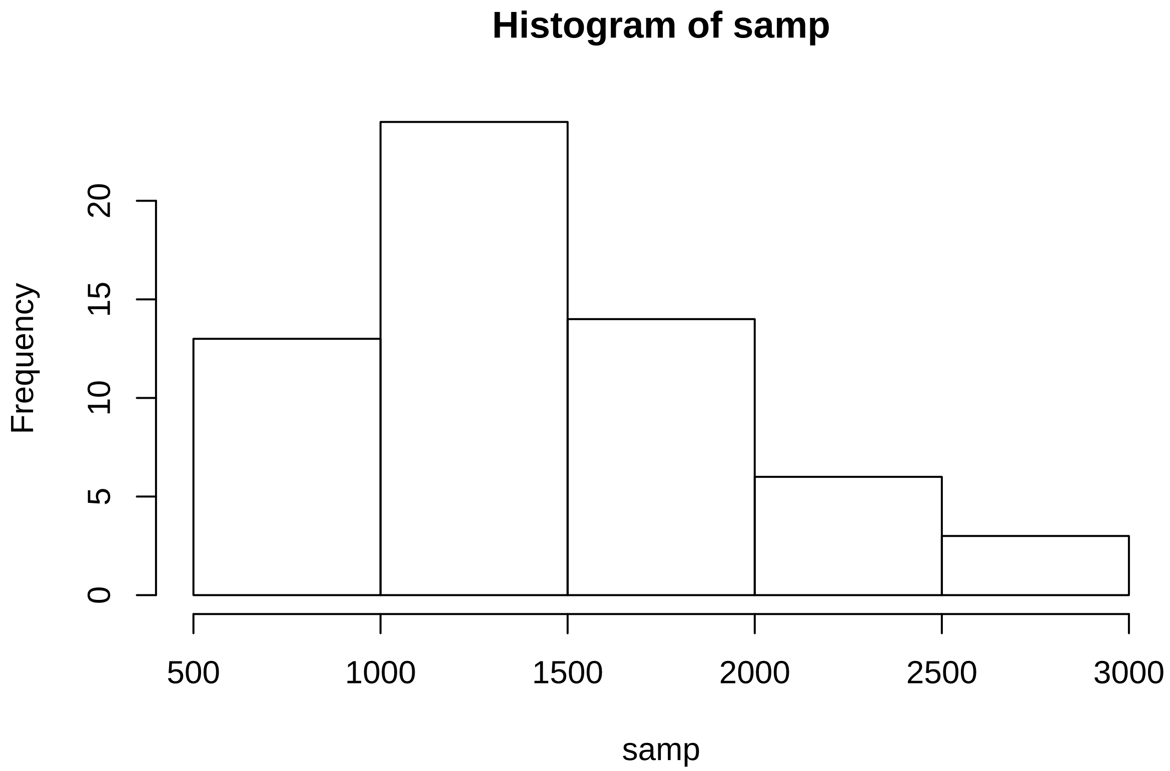
```r
load("more/ames.RData")
```

In this lab we'll start with a simple random sample of size 60 from the population. Specifically, this is a simple random sample of size 60. Note that the data set has information on many housing variables, but for the first portion of the lab we'll focus on the size of the house, represented by the variable `Gr.Liv.Area`.

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

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.

```r
hist(samp)
```



The distribution of this sample is right skewed. It appears that the "typical" size within the sample is between 1000 and 1500 square feet. By "typical", I utilized the distribution to indicate the most frequent sample means, which fall between these two values. To check, I took the mean of the sample and confirmed it to be at 1450 square feet.

```
mean(samp)
```

```
## [1] 1450.133
```

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

I would not expect another student's distribtution to be identical to mine. Since the low sample size is quite small, there will be a large amount of variability in which values make up my sample and the other student's sample.

Confidence intervals

One of the most common ways to describe the typical or central value of a distribution is to use the mean. In this case we can calculate the mean of the sample using,

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

Return for a moment to the question that first motivated this lab: based on this sample, what can we infer about the population? Based only on this single sample, the best estimate of the average living area of houses sold in Ames would be the sample mean, usually denoted as $\bar{x}$ (here we're calling it `sample_mean`). That serves as a good *point estimate* but it would be useful to also communicate how uncertain we are of that estimate. This can be captured by using a *confidence interval*.

We can calculate a 95% confidence interval for a sample mean by adding and subtracting 1.96 standard errors to the point estimate (See Section 4.2.3 if you are unfamiliar with this formula).

```
se <- sd(samp) / sqrt(60)
lower <- sample_mean - 1.96 * se
upper <- sample_mean + 1.96 * se
c(lower, upper)
```

```
## [1] 1350.985 1539.048
```

This is an important inference that we've just made: even though we don't know what the full population looks like, we're 95% confident that the true average size of houses in Ames lies between the values *lower* and *upper*. There are a few conditions that must be met for this interval to be valid.

3. For the confidence interval to be valid, the sample mean must be normally distributed and have standard error $s/\sqrt{n}$. What conditions must be met for this to be true?

The conditions that must be met in order for this to be true are:

(1) the sample size must be at least 30

(2) the distribution must not be skewed (normally distributed)

(3) each observation must be independent from one another

Confidence levels

4. What does "95% confidence" mean? If you're not sure, see Section 4.2.2.

95% confidence means that there is a 95% chance that the population mean falls within the lower and upper boundary that we determined from the point estimate. In a normal distribution, we can state that we are 95% confident that an interval that encompasses 1.96 standard errors from the sample proportion will capture the population proportion.

In this case we have the luxury of knowing the true population mean since we have data on the entire population. This value can be calculated using the following command:

```
mean(population)
```

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

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?

Yes, by referring back to the 95% confidence interval that we captured above, with a lower bound of 1350.98 and an upper bound of 1539.05, our population mean of 1499.69 falls within this confidence interval. I'm not working on this lab in a classroom, but I can suspect that my neighbor's interval will most likely capture this value as well.

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'm not working on this lab in a classroom, but I'd expect that 95% of the intervals would capture the true population mean. Since some point estimates by classmates (although very uncommon) will be more than 1.96 standard errors from the parameter of interest, there may be a few classmates that have intervals that do not capture the true population mean.

Using R, we're going to recreate many samples to learn more about how sample means and confidence intervals vary from one sample to another. *Loops* come in handy here (If you are unfamiliar with loops, review the Sampling Distribution Lab).

Here is the rough outline:

- Obtain a random sample.
- Calculate and store the sample's mean and standard deviation.
- Repeat steps (1) and (2) 50 times.
- Use these stored statistics to calculate many confidence intervals.

But before we do all of this, we need to first create empty vectors where we can save the means and standard deviations that will be calculated from each sample. And while we're at it, let's also store the desired sample size as `n`.

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

Now we're ready for the loop where we calculate the means and standard deviations of 50 random samples.

```
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
}
```

Lastly, we construct the confidence intervals.

```
lower_vector <- samp_mean - 1.96 * samp_sd / sqrt(n)
upper_vector <- samp_mean + 1.96 * samp_sd / sqrt(n)
```

Lower bounds of these 50 confidence intervals are stored in `lower_vector`, and the upper bounds are in `upper_vector`. Let's view the first interval.

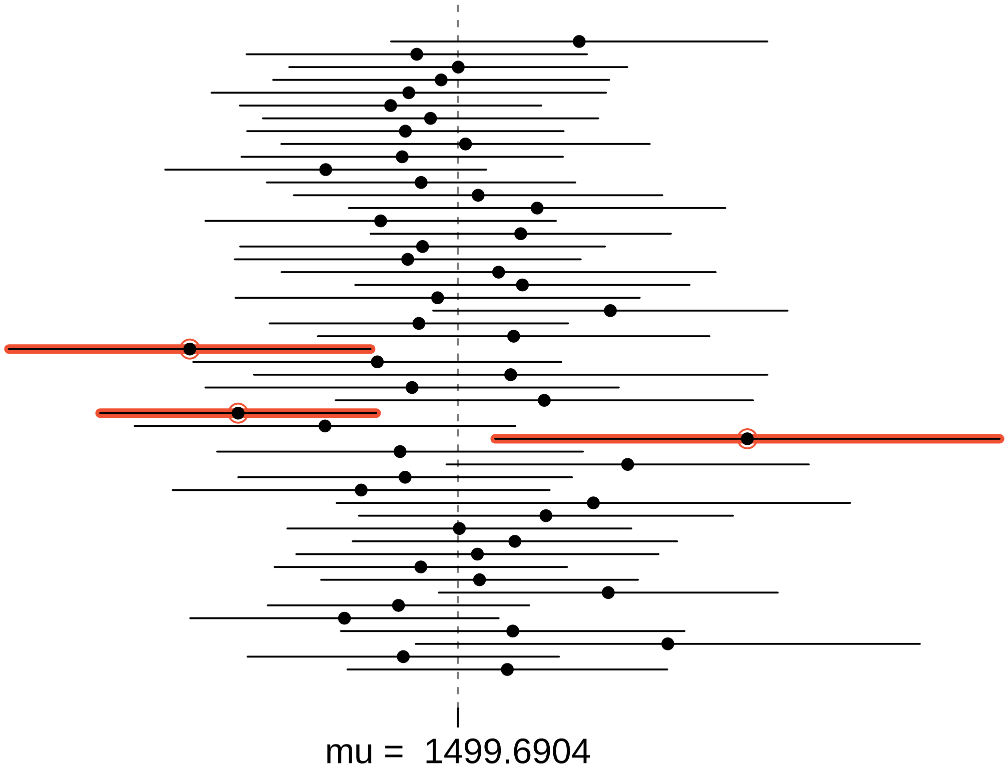
```
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1422.313 1646.254
```

On your own

- 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))
```



It appears that 47 out of 50 confidence intervals and point estimates from the simulation include the true population mean (94%). This proportion is not exactly equal to the confidence level. This is the case because just as observations naturally occur more than 1.96 standard deviations from the mean, some point estimates will be more than 1.96 standard errors from the parameter of interest. The confidence interval provides a logical range of values, and although we may say that other values are not very common, it doesn't mean that they aren't impossible. Therefore, this proportion of 94% is slightly lower than the 95% confidence interval.

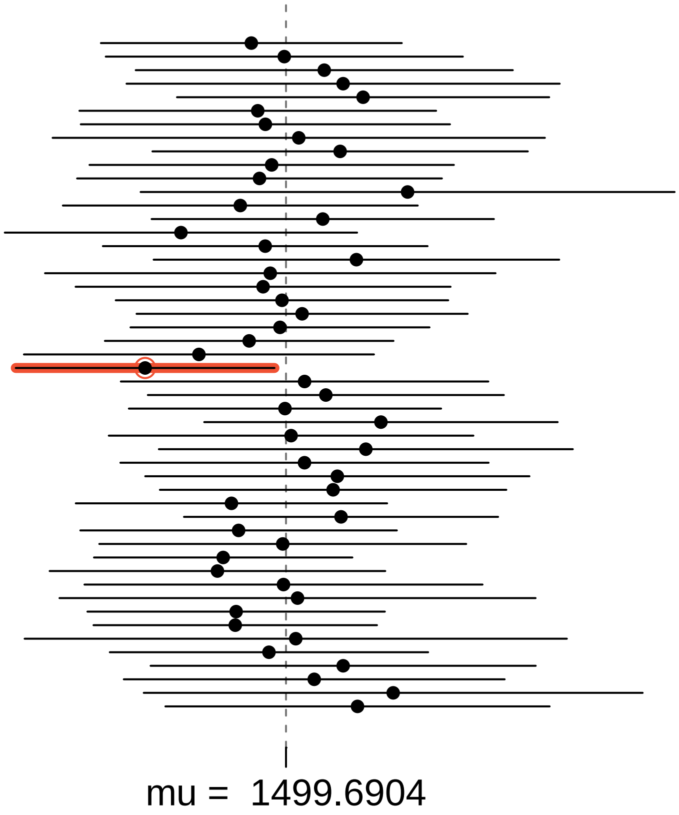
- Pick a confidence level of your choosing, provided it is not 95%. What is the appropriate critical value?

Since it is another common confidence level used throughout the textbook, I'll use a 99% confidence interval. The critical value for this confidence interval is $+/-2.85 * SE$.

- 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?

```
# changed the confidence level to +/- 2.85
lower_vector <- samp_mean - 2.85 * samp_sd / sqrt(n)
upper_vector <- samp_mean + 2.85 * samp_sd / sqrt(n)
```

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



It looks like 49 out of 50 confidence intervals and point estimates include the true population mean (98%). This is slightly below the confidence interval that was set at 99%.