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.

```
setwd("C:/Users/Hammer/Documents/Lab4b")
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`.

```
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.

We can describe the distribution of the sample using the output of R's **summary** and **sd** functions and via a histogram plot. **Please note that the results we obtain via these functions will vary each time we run the R code that generates the sample. Therefore, the specific results discussed here are representative solely of the results obtained at the time of this writing.** :

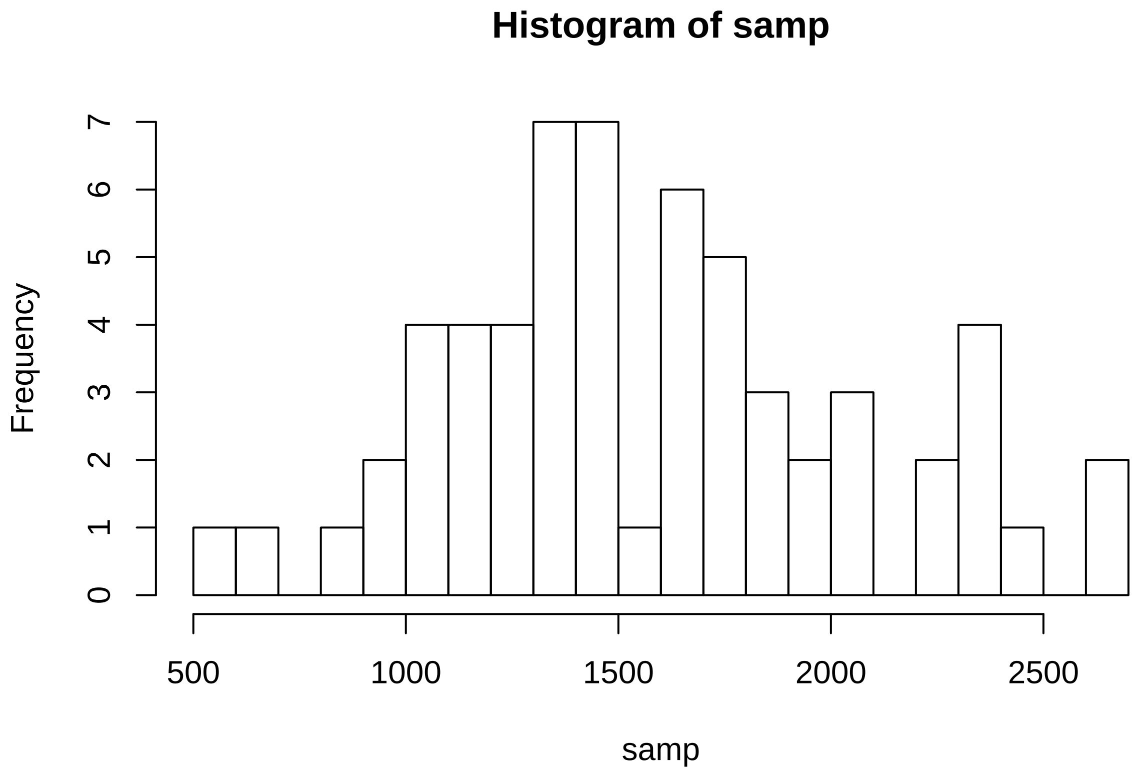
```
summary(samp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     520    1272    1460    1583    1842    2698
```

```
sd(samp)
```

```
## [1] 485.9648
```

```
hist(samp, breaks=20)
```



The distribution of the sample has a mean of 1519, a standard deviation of 426.331, and a median of 1496. It is right skewed as evidenced by the fact that the value of the mean exceeds the value of the median by a meaningful amount. Further evidence of right skew is obvious in the appearance of the histogram. Based on this one sample our point estimate of the "typical" size of a house in Ames, Iowa is 1519 square feet.

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 expect another student's distribution to be similar but not identical to mine since each of us will likely have selected different elements from the population via sampling.

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] 1459.917 1705.849
```

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?

From the textbook we are told:

- The sample observations are independent.
- The sample size is n > 30.
- If sampling without replacement, the sample size n must be < 10% of the population
- The population distribution is not strongly skewed.

Confidence levels

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

If we took many samples and built a confidence interval from each sample using the central limit theorem, then about 95% of those sample confidence intervals will contain the actual population mean.

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?

The confidence interval we calculated earlier was (1300, 1516). Since the actual population mean is 1500, it appears that the confidence interval does capture the true average size of houses in Ames.

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 expect 95% of my fellow students' confidence intervals to capture the true population mean since the conditions required for a valid confidence interval (see Exercise 3 above) have been met.

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`.

```r
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.

```r
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.

```r
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.

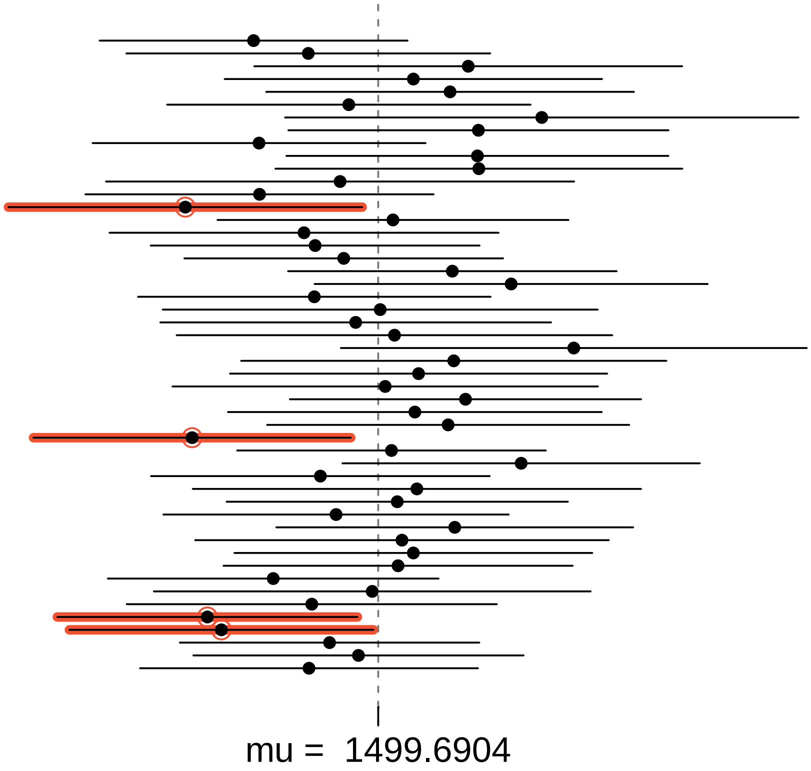
```r
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1338.858 1567.042
```

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.

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



What proportion contain the mean? 46/50, or 0.92 of the confidence intervals include the true population mean:

```
46/50
```

```
## [1] 0.92
```

Is this exactly equal to the confidence interval?
NO - it is less than 95% confidence level. The difference could be due to the fact that the distribution is not normal; it is right-skewed. The right-skewness results in a wider standard deviation than would a be attributed to the population if it were normally distributed.

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

If we select a confidence level of 98% the critical value would be 2.33.

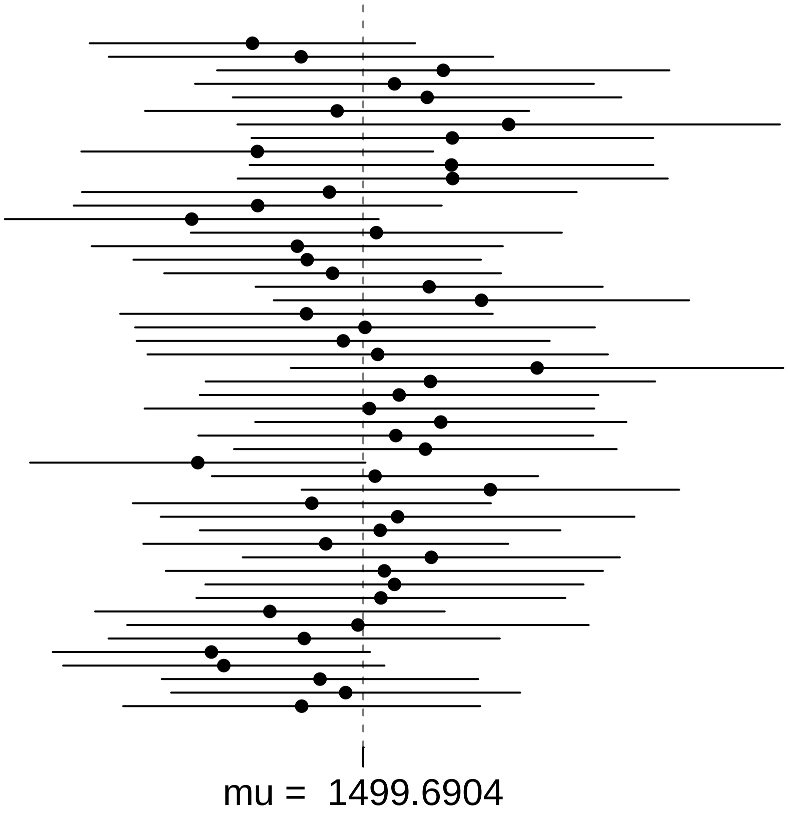
- Calculate 50 confidence intervals at the confidence level you chose in the previous question. You do not need to obtain new samples, simply calculate new intervals based on the sample means and standard deviations you have already collected. Using the `plot_ci` function, plot all intervals and calculate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

```
# construct the confidence intervals for the 50 random samples at a 98% conf level

lower_vector <- samp_mean - 2.33 * samp_sd / sqrt(n)
```

```
upper_vector <- samp_mean + 2.33 * samp_sd / sqrt(n)

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

```
## [1] 1317.321 1588.579
```

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



The proportion of intervals that include the population mean is 49/50, or 0.98, which matches the selected confidence level of 98%.