

Foundations for statistical inference - Confidence intervals

Ilya Kats

March 12, 2017

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.

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

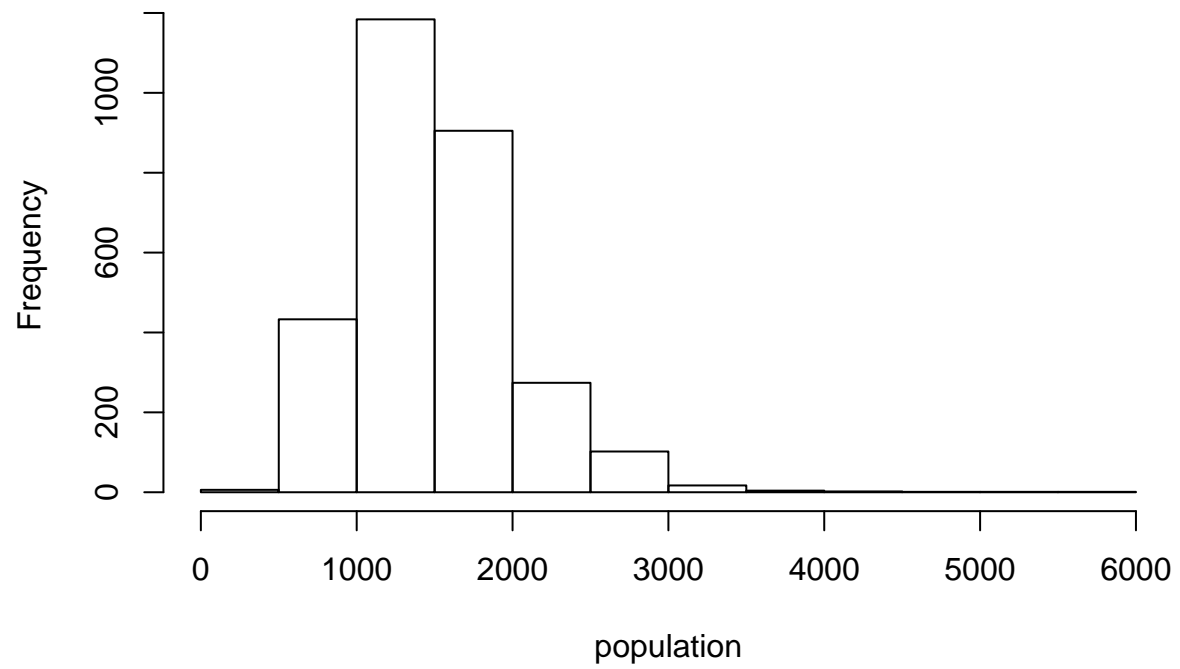
In this lab we'll start with a simple random sample of size 60 from the population. Specifically, this is a simple random sample of size 60. Note that the data set has information on many housing variables, but for the first portion of the lab we'll focus on the size of the house, represented by the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

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

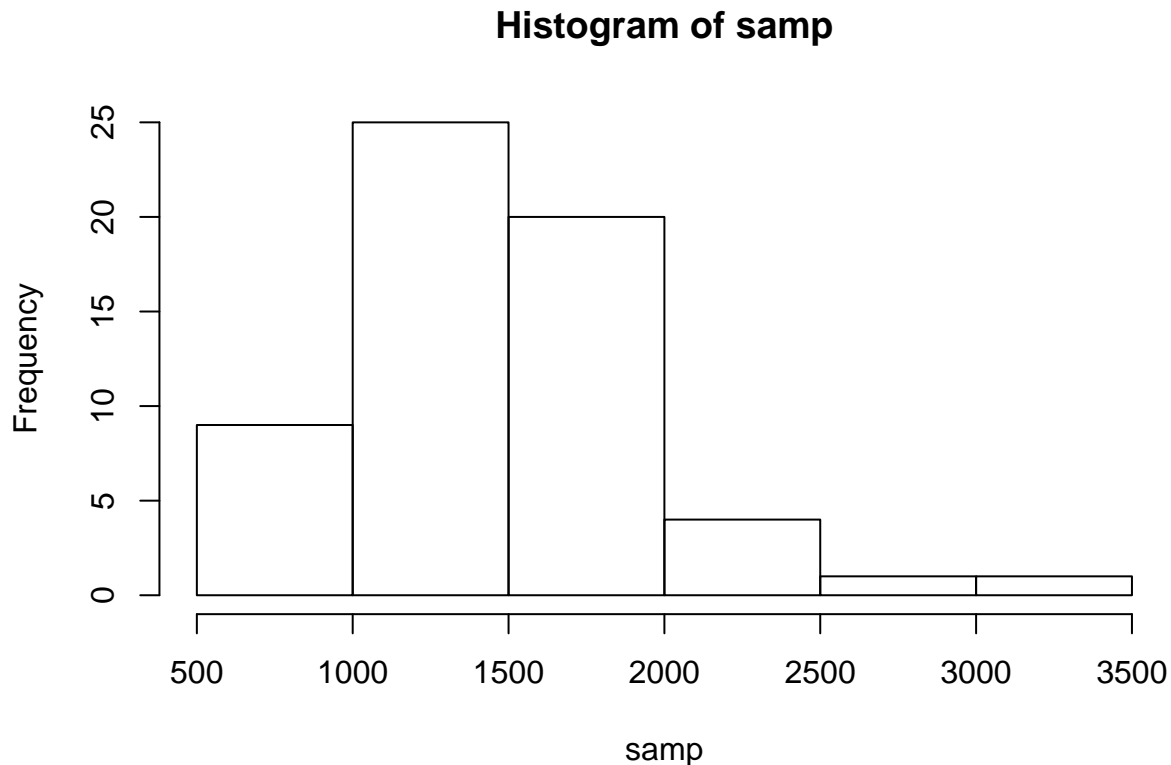
1. Describe the distribution of your sample. What would you say is the "typical" size within your sample? Also state precisely what you interpreted "typical" to mean.

```
hist(population)
```

Histogram of population



```
hist(samp)
```



The population distribution is unimodal with a slight right-skew, so the distribution of the sample should be similarly unimodal with a weak right-skew. Based on the sample I would expect the typical size to be around 1500 square feet. The typical size is the mean of the sample which is a point estimate of the true population mean.

-
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 distribution to be identical, but I would expect it to be similar. The distribution of a sample of size 60 is representative of the population distribution because the sample is random. However, since it is only a sample there will be variation between samples (and at rare occasions differences can be significant).

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] 1375.239 1607.695
```

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?

Conditions are as follows: sample observations must be independent, sample size should be equal to or more than 30 and population distribution should not be strongly skewed.

Confidence levels

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

It means that we are 95% confident that the population mean is between the lower and upper boundary we determined from the sample mean (point estimate). In other words if we continue generating samples and calculating boundaries, 95% of those boundaries will contain the true 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?

Confidence interval does capture the true population mean.

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.
-

As described in answer 4, I would expect 95% of intervals to capture the true population mean. In normal distribution 95% of observations are within the 1.96 times of standard deviation (standard error).

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.

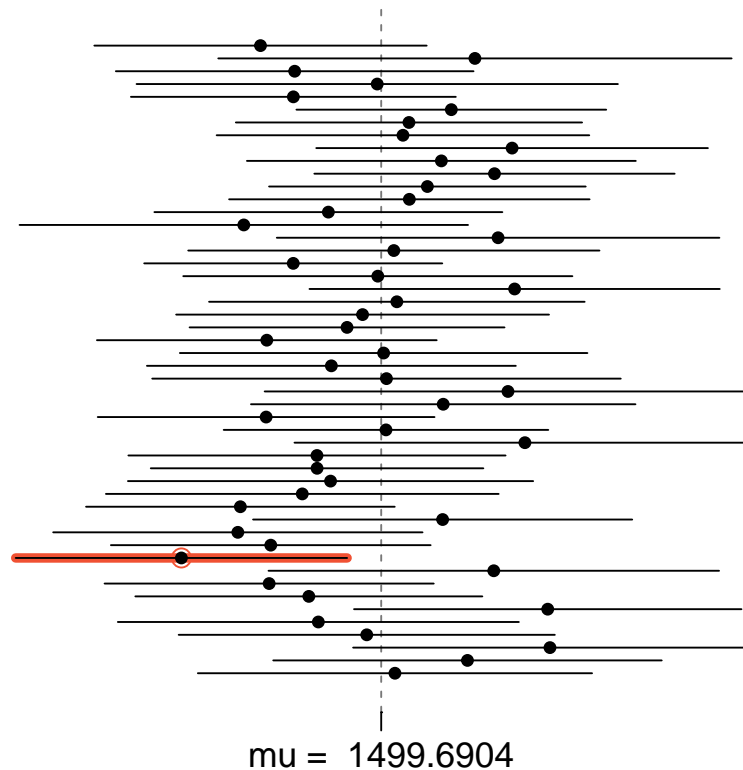
```
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1379.060 1638.407
```

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))
```



```
pop_mean <- mean(population)
outside_intervals <- length(lower_vector[lower_vector > pop_mean]) + length(upper_vector[upper_vect
```

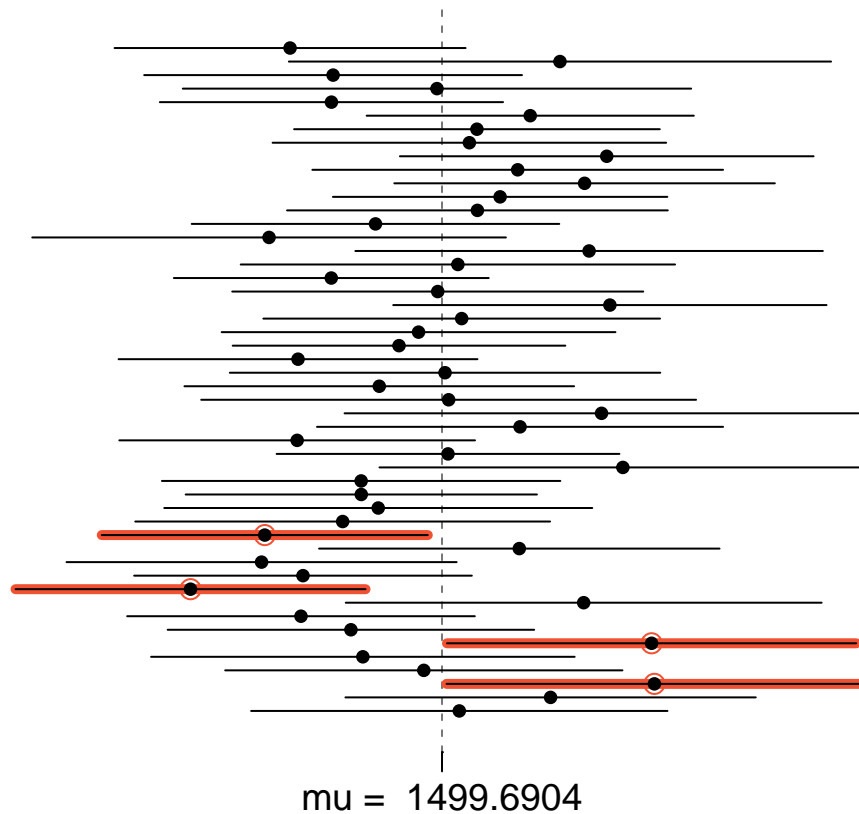
Out of fifty confidence intervals, there are 1 intervals that do not include the true population mean. That is 2% of the intervals which is similar to the confidence level (exact value is not possible with 50 intervals since it will require 2.5 intervals to be outside of the population mean).

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

Let's pick a confidential level of 90%. The appropriate critical value is 1.6448536.

- 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?

```
lower_vector90 <- samp_mean - qnorm(0.95) * samp_sd / sqrt(n)
upper_vector90 <- samp_mean + qnorm(0.95) * samp_sd / sqrt(n)
plot_ci(lower_vector90, upper_vector90, mean(population))
```



```
outside_intervals90 <- length(lower_vector90[lower_vector90 > pop_mean]) + length(upper_vector90[upper_vector90 > pop_mean])
```

There are 4 intervals that do not contain the true population mean. That is 8% of the intervals or very close to the 90% confidence interval. Small variations are acceptable.

This is a product of OpenIntro that is released under a Creative Commons Attribution-ShareAlike 3.0 Unported. This lab was written for OpenIntro by Andrew Bray and Mine Çetinkaya-Rundel.