

DATA606_Lab5b_RJM

RJM

2020-01-07

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

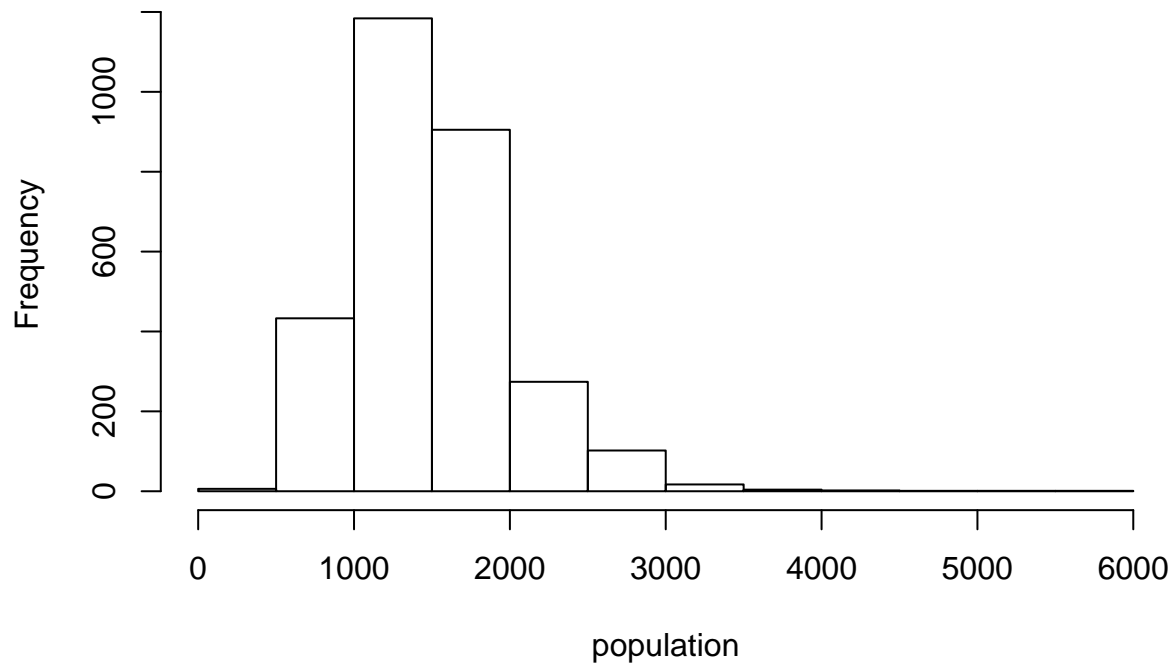
In the previous lab, "Sampling Distributions", we looked at the population data of houses from Ames, Iowa. Let's start by loading that data set.

```
download.file("http://www.openintro.org/stat/data/ames.RData",destfile="ames.RData")
load("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
set.seed(123)
hist(population)
```

Histogram of population



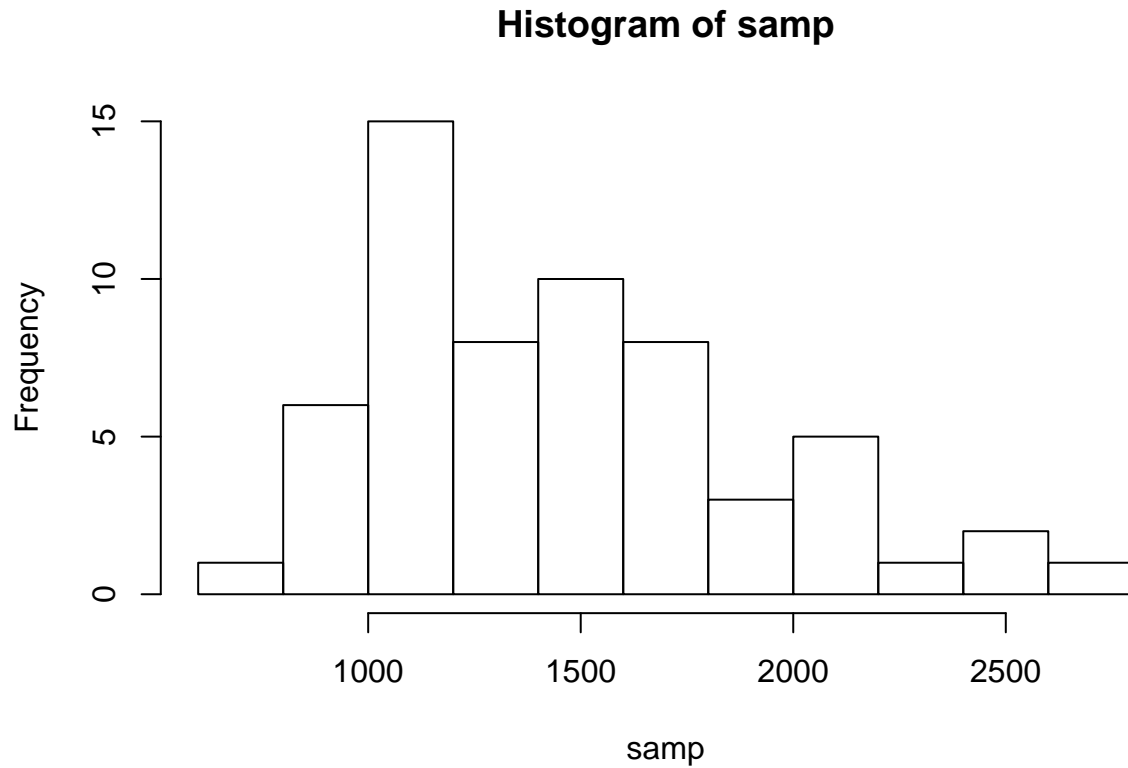
```
summary(population)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max. \n##      334   1126   1442   1500   1743   5642
```

```
sd(population)
```

```
## [1] 505.5089
```

```
set.seed(123)\nsamp <- sample(population, 60)\nhist(samp)
```



```
summary(samp)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      767   1118   1398   1462   1679   2610
```

```
sd(samp)
```

```
## [1] 446.8552
```

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.

The distribution seems to be unimodal with a slight right skew. A typical size would be between 1,000 - 2,000 sq ft with a majority being around 1,500 sq ft. The typical here would mean a size that represents the most prevalent in the sample.

2. Would you expect another student’s distribution to be identical to yours? Would you expect it to be similar? Why or why not? Another student’s distribution will be identical only if they chose the same seed value as mine. However, it will be similar as the same number of random samples are drawn from the same population.

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] 1348.513 1574.653
```

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?

In order for this to be true, the following conditions must hold: 1. Independence of the sample observations. 2. The sample size is large enough (30 or more). 3. The distribution of population is not skewed. It is difficult to establish a skew but a slight one is tolerable.

Confidence levels

4. What does "95% confidence" mean? If you're not sure, see Section 4.2.2. A 95% confidence means a range of values that we are 95% confident to contain the population mean. The bigger the sample gets, narrower this range becomes.

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 has a range of 1415.997 to 1651.436, which captures the true average size of houses in Ames dataset. I am not in a classroom.
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 we have set a 95% significant level (1.96 z-score), it is expected that 95% of the intervals will have the true population mean. I am not in a classroom.

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`.

```
set.seed(123)
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] 1348.513 1574.653
```

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.

Looking at the plot below, we can see that there is only one interval where the true population mean is not included. This means that 98% of the confidence intervals include the true population mean. This shows that we have more intervals containing the true population mean than 95%. It is due to the fact that we are using random samples and we can have intervals containing the true population mean hovering above and below the 95% mark without much variance.

```

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

<!-- -->

```

- Pick a confidence level of your choosing, provided it is not 95%. What is the appropriate critical value?
I have picked a confidence level of 90% with a critical value of 1.64.
- 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?

At 90% confidence interval, we get 96% (48 out of 50) of intervals contain the true population mean. This is still higher than the expected 90%, but lower than the 98% from the last test where we had the confidence interval set at 95%.

```

lower_vector <- samp_mean - 1.64 * samp_sd / sqrt(n)
upper_vector <- samp_mean + 1.64 * samp_sd / sqrt(n)

plot_ci(lower_vector, upper_vector, mean(population))

```

