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.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

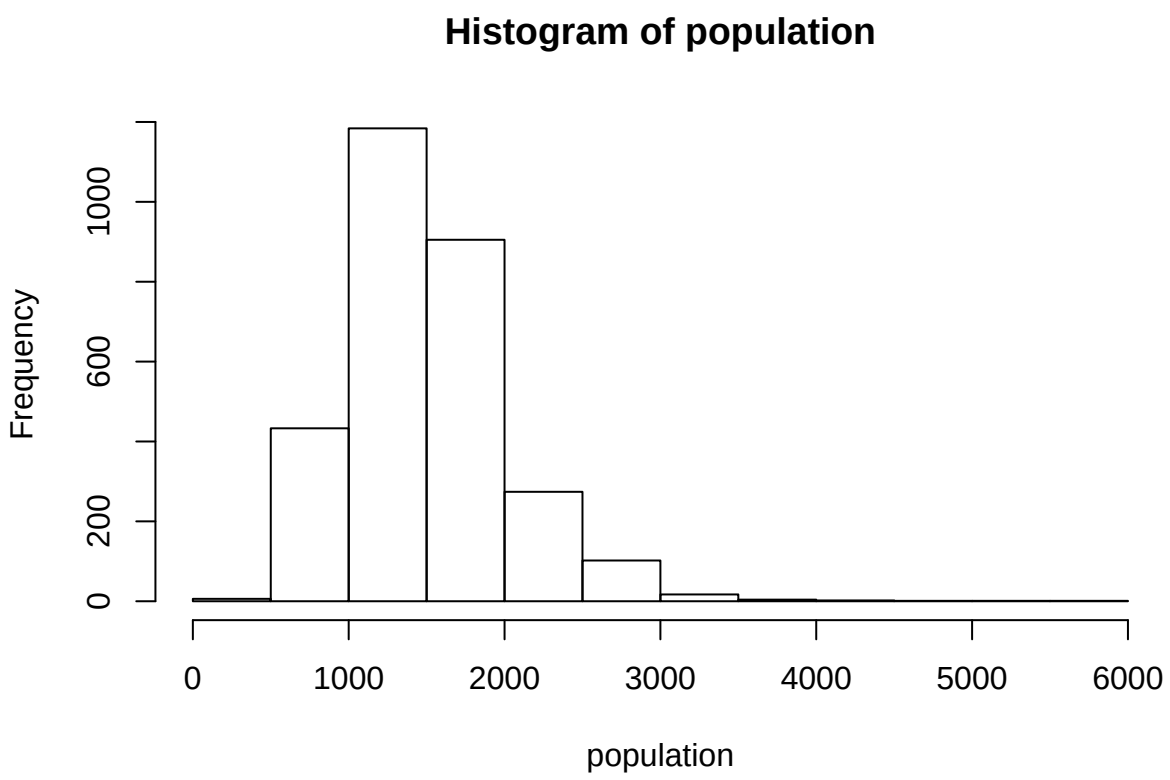
```
load("more/ames.RData")
```

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`.

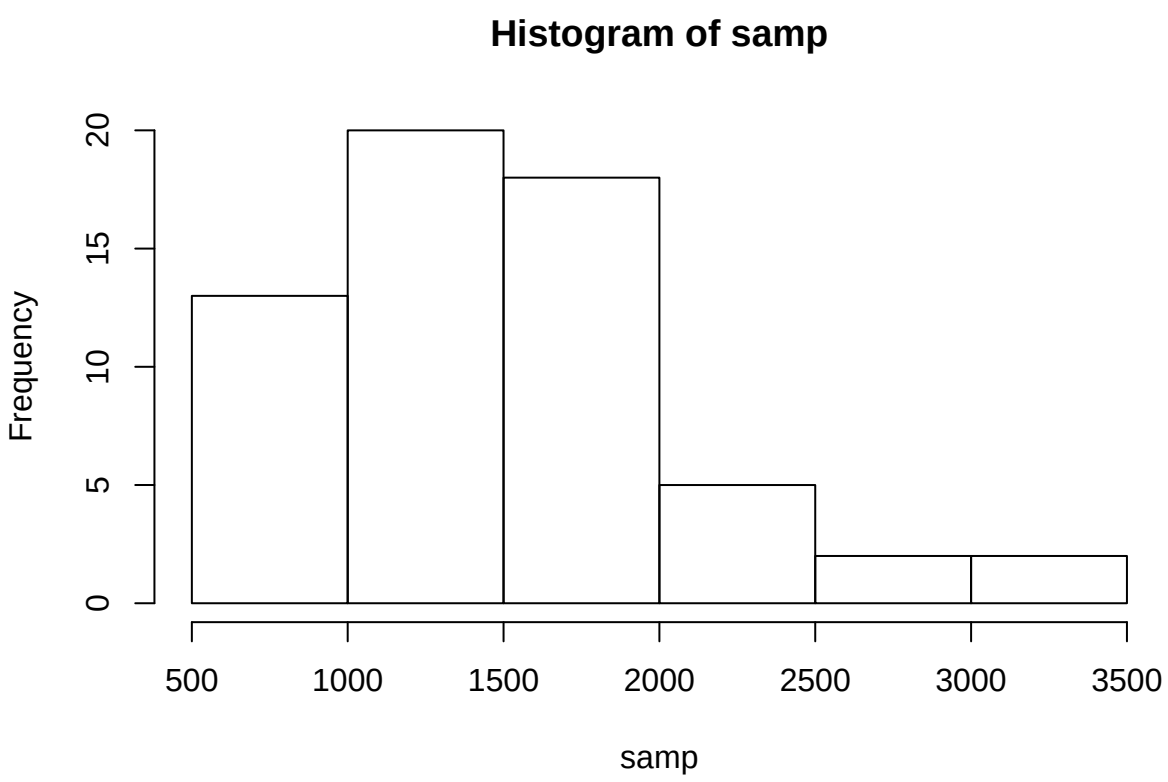
```
population <- ames$Gr.Liv.Area
samp <- sample(population, 60)
```

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
Frequency
0
200
600
1000
0
1000
2000
3000
4000
5000
6000
population

```
hist(samp)
```



The distribution of the sample is normally distributed. Typical size means, the minimum sample size for analyzing the population distribution. Although the popuation distribution is right skewed, if we fetch decent sample size of 30, the sample will be normally distributed. And the each sample should be independent of another 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?

Yes, most of the student's distribution will be identical or normally distributed. This is because of centeral theorem. It happens only if we have a decent sample (more than 30).

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)
mean(population)
```

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

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] 1360.48 1640.72
```

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?

1. Each sample should be independent. It should have equal probability of fetching. 2. If the population distribution is skewed, then the sample size should be atleast 30. So that we can use z distribution. If sample size is less than 30, then we an use the T-distribution(given that the population is normally distributed) 3. Larger the sample size, more lenient we can be.

Confidence levels

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

** We can define it as multiple different ways.

1. 95% confidence means that we are 95% confident that the taken samples will contain the point estimate in population.
2. The mean of the sample with two standard deviations (95% confidence interval) will contain the population mean or expected value of population. **

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?

```
samp <- sample(population, 60)

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

```
## [1] 1385.836 1615.364
```

```
mean(population)
```

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

If we run the above code multiple times, we might see that the confidence interval varies for different sample sizes. But in all the cases true mean falls in that specific confidence interval. This shows that the population mean will be in the confidence interval.

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.

```
samp <- select(head(arrange(ames,Gr.Liv.Area),90),Gr.Liv.Area)

samp <- sample(samp$Gr.Liv.Area, 60,replace = FALSE)

sample_mean <- mean(samp)
se <- sd(samp) / sqrt(60)
lower <- sample_mean - 2.96 * se
upper <- sample_mean + 2.96 * se
c(lower, upper)
```

```
## [1] 677.1135 752.6865
```

```
mean(population)
```

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

We expect that the 95% of intervals will capture the true population mean. But still there are chances of making type I error.

For example, above code is a test to find out some rare scenario. We arrange and fetch the 90 observations into a data frame(purposefully used the lower number of data). If the sample contains all the smaller values and then we figure out the confidence interval, we are not actually capturing the true population mean. Depends on confidence interval, we can say that we are 95% confident that the samples will contain 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.

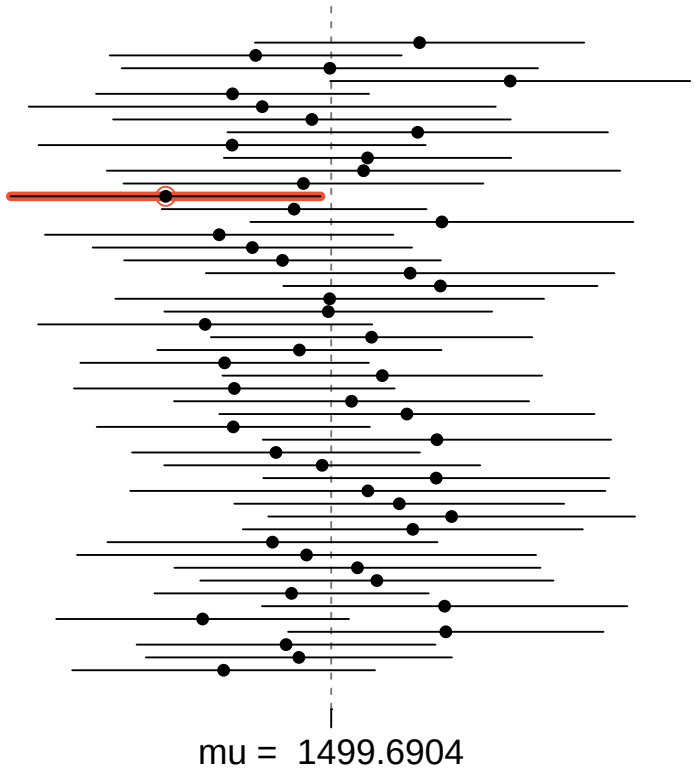
```
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1303.080 1532.953
```

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))
  ```




```
interval <- data.frame(lower_vector,upper_vector)
truepopuation <- mean(population)

interval <- mutate(interval,intruevalue=lower_vector<truepopuation & upper_vector<truepopuation)

filter(interval,intruevalue==TRUE)
```

```
##   lower_vector upper_vector intruevalue
## 1     1256.351     1491.649        TRUE
```

Above code shows the most of the values are in the 95% confidence interval. Only one value is not in that interval. It may also be more than 95% confidence interval.

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

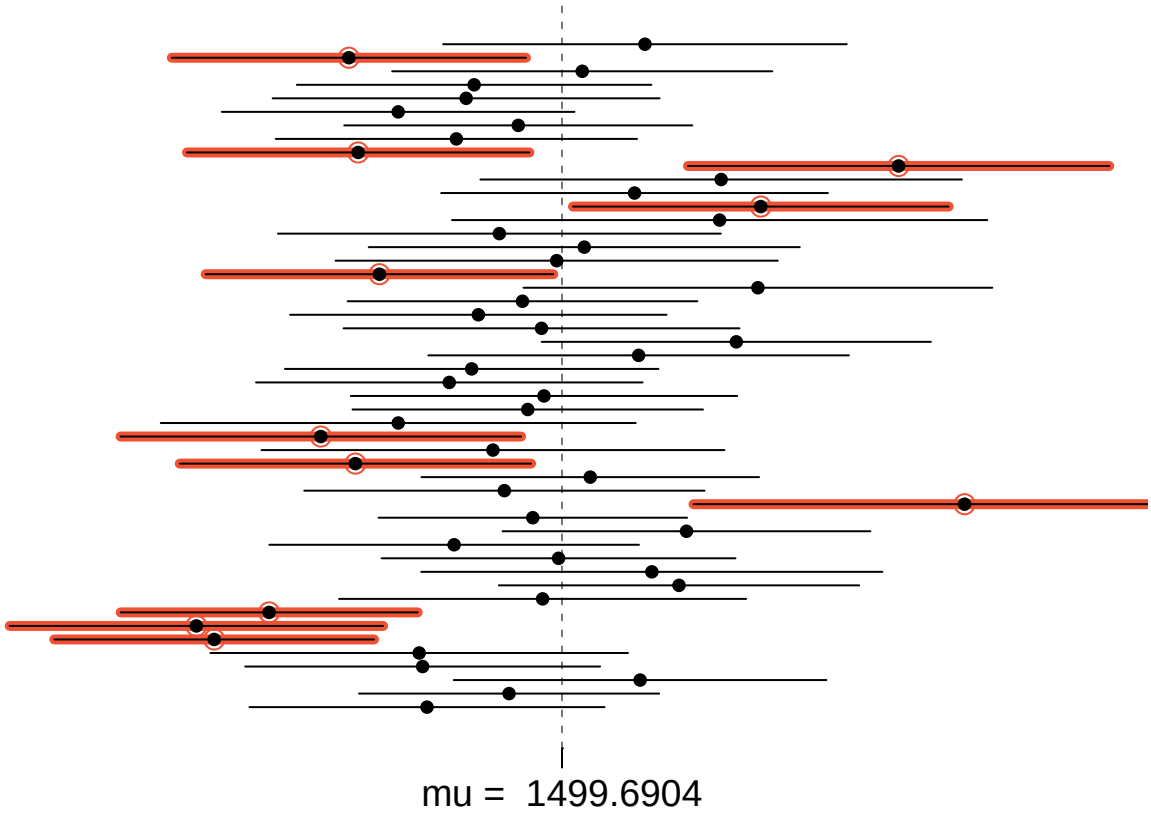
```
for(i in 1:50){
  samp <- sample(population, n) # obtain a sample of size n = 60 from the population
  samp_mean[i] <- mean(samp)    # save sample mean in ith element of samp_mean
  samp_sd[i] <- sd(samp)        # save sample sd in ith element of samp_sd
}


lower_vector <- samp_mean - qnorm(.90) * samp_sd / sqrt(n)
upper_vector <- samp_mean + qnorm(.90) * samp_sd / sqrt(n)
sd_vector <- samp_sd
interval <- data.frame(lower_vector,upper_vector,sd_vector)
```

```
interval <- mutate(interval,intruevalue=lower_vector<truepopuation & upper_vector<truepopuation,low
```

Above code shows the 90% confidence interval. It also calculates the critical value by calcuating the z-score value with upper and lower limit.

- 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?

```
plot_ci(interval$lower_vector,interval$upper_vector,m = mean(population))
```



Above chart shows the confidence interval of 90%. About ~10% of 60 samples does not have the true mean of the population