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.

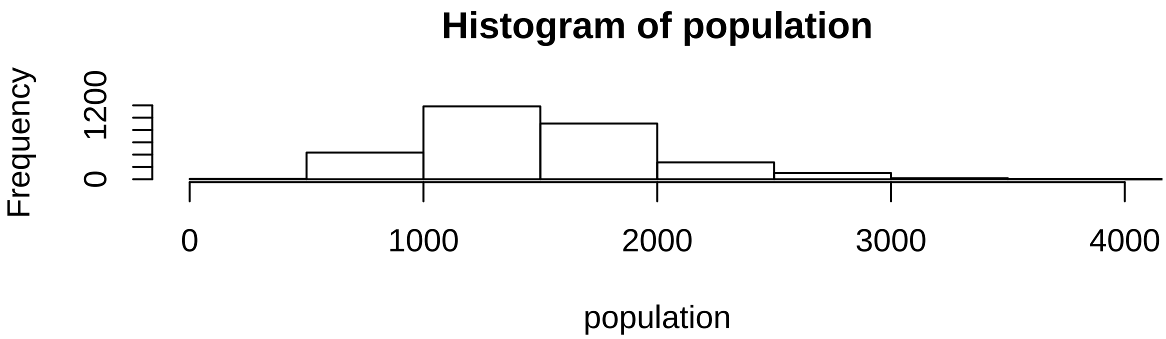
```
load("more/ames.RData")
```

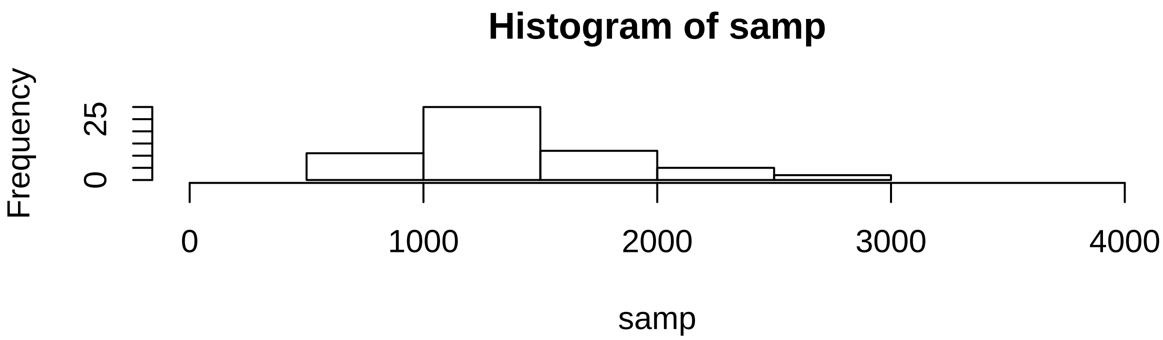
In this lab we'll start with a simple random sample of size 60 from the population. Specifically, this is a simple random sample of size 60. Note that the data set has information on many housing variables, but for the first portion of the lab we'll focus on the size of the house, represented by the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

```
population <- ames$Gr.Liv.Area
samp <- sample(population, 60)
par(mfrow=c(2,1))
str(population)
```

```
##  int [1:2930] 1656 896 1329 2110 1629 1604 1338 1280 1616 1804 ...
```

```
hist(population, xlim=c(0,4000))
hist(samp, xlim=c(0,4000))
```





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.

```
summary(population)
```

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

```
summary(samp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     694    1083    1367    1427    1710    2683
```

The distribution of the sample looks like an unimodal normal distribution. I would consider the median value of the sample to be typical size of the houses. This value changes everytime we re-sample the population. 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 will expect most student's distributions to be different than mine as they sample 60 values from 2930 possible values.*

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
confidence.interval <- c(lower, upper)
confidence.interval
```

```
## [1] 1308.358 1546.142
```

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? *Normal distributions are symmetric, unimodal, and asymptotic, and the mean, median, and mode are all equal*

Confidence levels

4. What does "95% confidence" mean? If you're not sure, see Section 4.2.2. *"95% confidence" means that in about 95 percent of the samples the true value of of the observation is contained in the confidence interval obtained from the sample.* 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
```

```
## [1] 1308.358 1546.142
```

The confidence interval (1361.823, 1599.810) calculated previously in this lab does indeed captures the mean value (1499.69) of the population.

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. *95% of the calculated intervals should 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)
head(lower_vector)
```

```
## [1] 1260.719 1371.205 1317.412 1414.704 1304.842 1316.624
```

```
head(upper_vector)
```

```
## [1] 1500.281 1676.929 1557.321 1666.596 1538.525 1523.610
```

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.

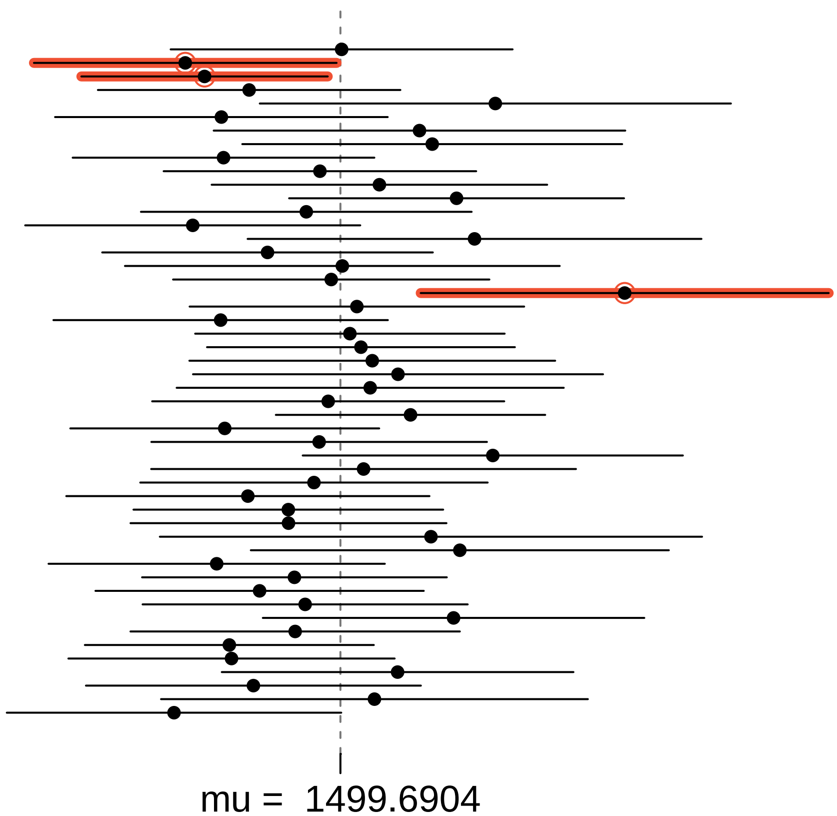
```
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1260.719 1500.281
```

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. *From the plot below we see that 49 (or 98%) of the calculated confidence intervals contained the true population mean. This proportion is higher than the chosen confidence value of 95%. This is OK because everytime we choose a sample population we get a new subset of values. 95% or more of those will contain the true population mean.*

```
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?

If we choose a confidence level of 90%, the critical value or z-score is 1.645

- 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?

Constructing the confidence intervals for 90% confidence level.

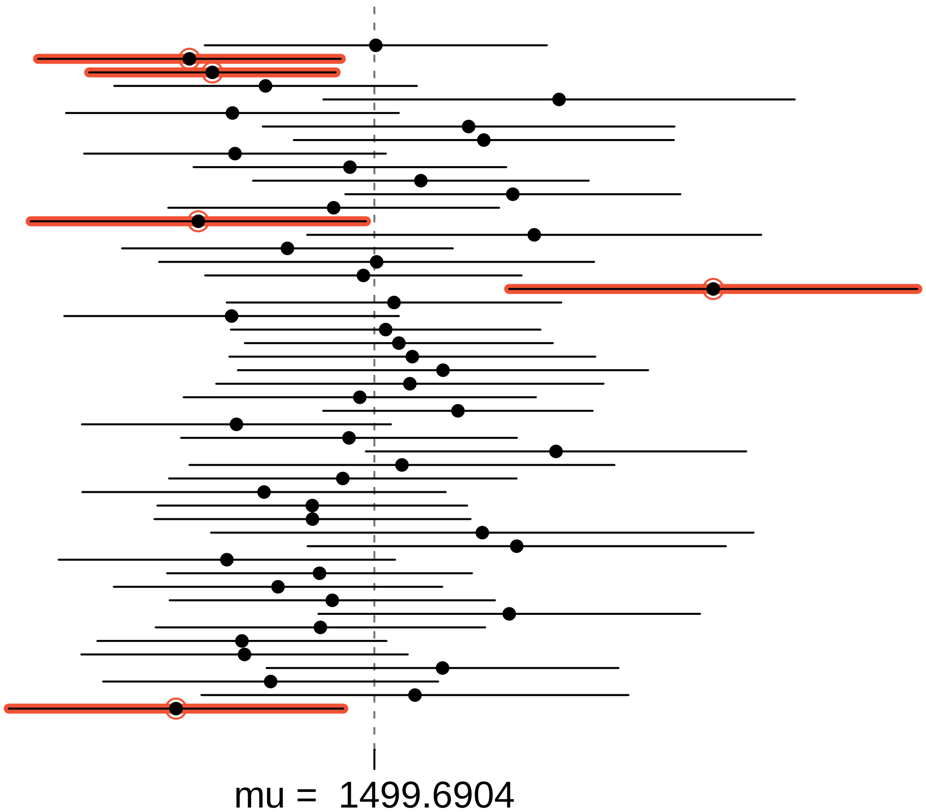
```
lower_vector.90 <- samp_mean - 1.645 * samp_sd / sqrt(n)
upper_vector.90 <- samp_mean + 1.645 * samp_sd / sqrt(n)
head(lower_vector.90)
```

```
## [1] 1279.969 1395.772 1336.690 1434.946 1323.620 1333.256
```

```r
head(upper_vector.90)
```

```
## [1] 1481.031 1652.362 1538.043 1646.354 1519.747 1506.977
```

```r
plot_ci(lower_vector.90, upper_vector.90, mean(population))
```



In the calculated range we see that with a chosen confidence interval o 90%, 48 of our interval ranges (96%) contain the true mean of the population.