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)
```

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.

Exercise 1 Answer:

```
hist(samp, breaks=40)
```

Histogram of samp
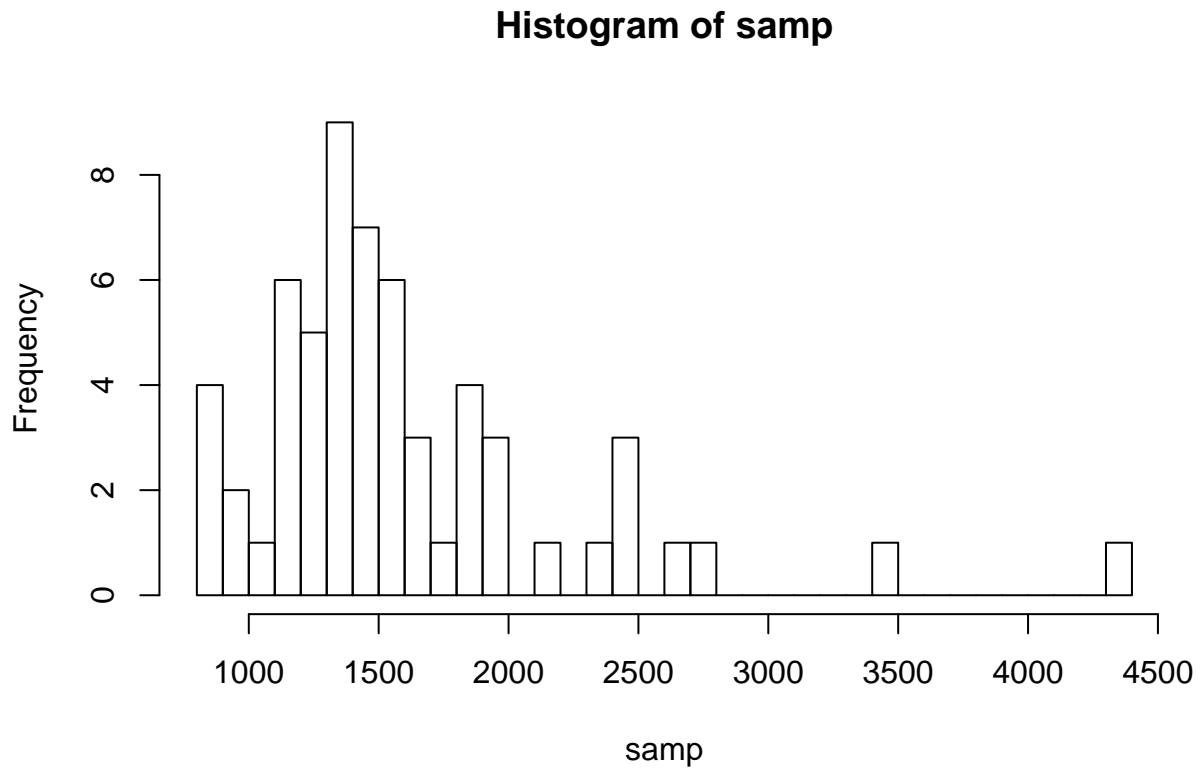

```
# I would describe the sample as unimodal around 1700 and fairly
# flat otherwise, with some extreme outliers on the high side.
summary(samp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     864    1219    1436    1604    1832    4316
```

```
# I would say the "typical" size is around 1600-1700 because the largest frequencies
# occur there, and the sum of the two highest bars is equal to or greater than
# the sum of the 3-4 bars on either side of the two highest bars.
```

2. Would you expect another student's distribution to be identical to yours? Would you expect it to be similar? Why or why not?

Exercise 2 Answer:
Another student's distribution would NOT be identical to mine. I would expect it to be somewhat similar, but it could vary considerably because we're all taking single samples and the results of my sample was far from perfectly normal, so I'd expect considerable variability in samples.

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] 1447.035 1761.965
```

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?

Exercise 3 Answer:

There are three conditions:
1. The sample observations must be independent
2. The sample size must be equal to or greater than 30
3. The population distribution must not be strongly skewed

Confidence levels

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

Exercise 4 Answer:

It means that if we took multiple samples, then approximately 95% of the sample means would be within the range set by the confidence interval based on the single sample and the point estimate.

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?

Exercise 5 Answer:

Not necessarily. A 95% confidence interval means there is still a 5% chance that the population mean does not fall into the interval. The same would be true for any classmate constructing a confidence interval from a single sample.

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.

Exercise 6 Answer:

I would expect approximately 95% of the confidence intervals to capture the true population mean, because the intervals created were 95% confidence intervals. I am not working on this lab 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.

```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] 1490.827 1752.007
```

4

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))
```


mu = 1499.6904
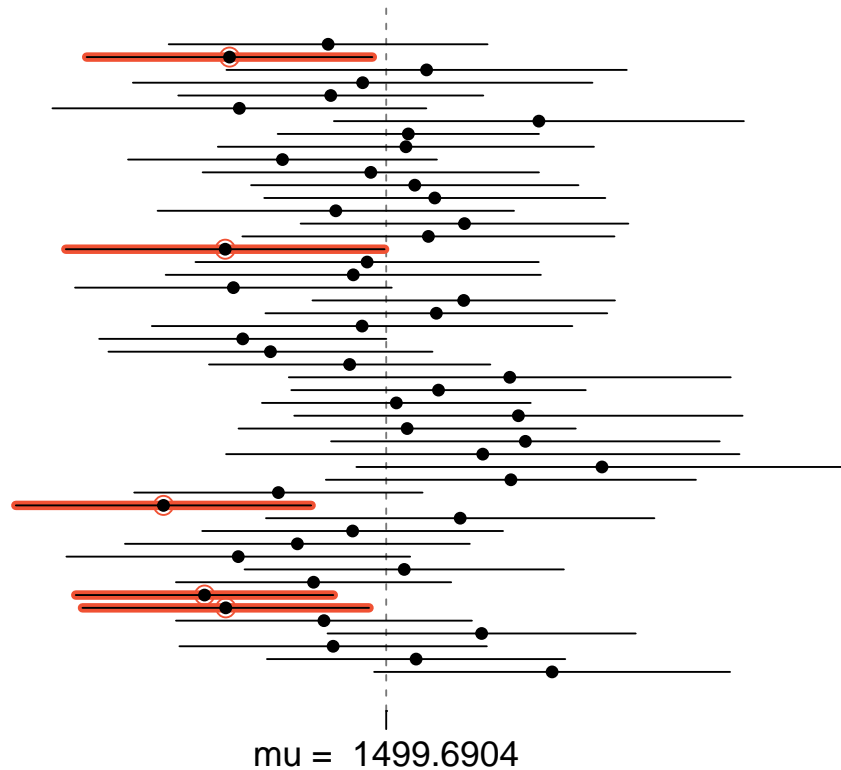
- Pick a confidence level of your choosing, provided it is not 95%. What is the appropriate critical value?

- 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?

Number 1 Answer:

There are 50 confidence intervals, and 46 (or 92%) of them include the true population mean. This proportion is not the same as the confidence interval (95%). That's because the confidence interval is approximate. The larger the set of samples we had, the more the proportion would move towards 95%.
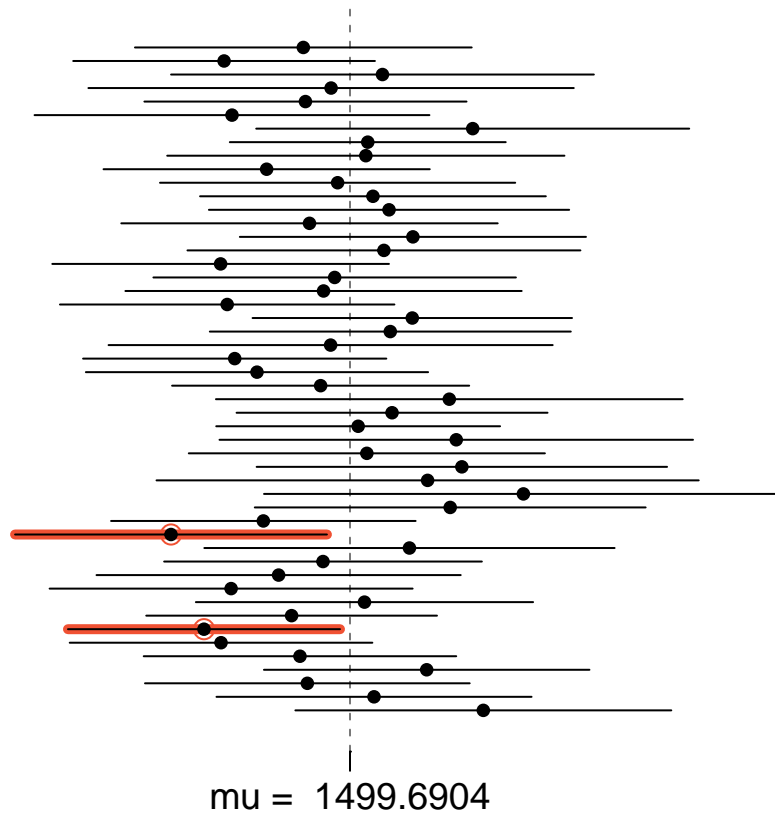
Number 2 Answer:

```
# Calculate critical value for 99% confidence interval
qnorm(0.995)
```

```
## [1] 2.575829
```

Number 3 Answer:

```
# Calculate 50 confidence intervals at a 99% confidence level
lower_vector <- samp_mean - qnorm(0.995) * samp_sd / sqrt(n)
upper_vector <- samp_mean + qnorm(0.995) * samp_sd / sqrt(n)
# Plot all intervals
plot_ci(lower_vector, upper_vector, mean(population))
```


mu = 1499.6904

The plot shows that 49 out of 50, or 98%, of the intervals include the true population mean. Though not exactly 99%, it is closer to the confidence interval than what we saw with the 95% confidence interval example for question number 1.

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.