Hao-4b

Bruce Hao

September 27, 2016

```
#setwd("C:/Users/bhao/Google Drive/CUNY/git/DATA606/Lab4b")
library(IS606)
library(dplyr)
library(ggplot2)
library(ggthemes)
set.seed(123)
load("more/ames.RData")
```

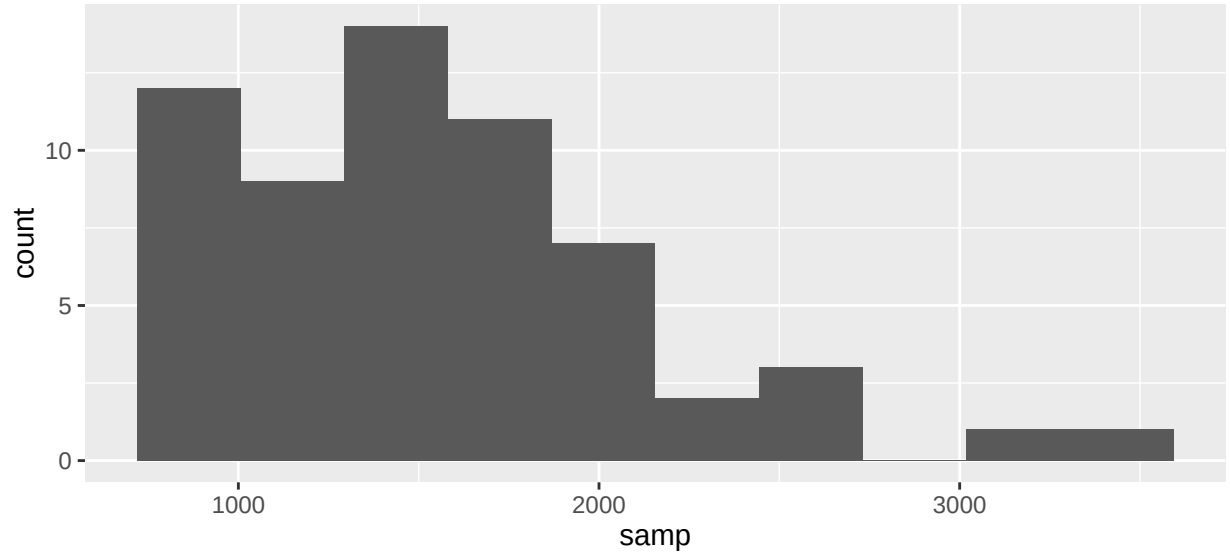
Exercise 1: Describe the distribution of your sample. What would you say is the "typical" size within your sample? Also state precisely what you interpreted "typical" to mean.

The sample has a mean of 1534 and median of 1413. Its distribution is right skewed with some outliers to the right. If I had to choose a 'typical' size from this sample, I would use the median of 1413 as the mean is more biased by the outliers.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     808    1061    1413    1534    1816    3395
```

```
qplot(samp, geom = 'histogram', bins = 10)
```



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

No, it would not be identical as the sample is a random sample, and so differences should be expected. Given the population size of 2930 and sample size of 60, I would expect that the results would be somewhat similar.

Exercise 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?

- Sufficient sample size - at least 30 independent observations
- Independence - usually, if the sample represents less than 10% of the population, the sample can be considered independent
- Little skew - if there is strong skew, a larger sample may be needed

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

If we took many samples and built confidence intervals, 95% of those intervals would contain the actual mean. Therefore, we can be roughly 95% confident that we have captured the true parameter within our confidence interval.

Excercise 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?

Yes.

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

```
## [1] 1390.046 1677.388
```

```
mean(population)
```

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

Exercise 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.

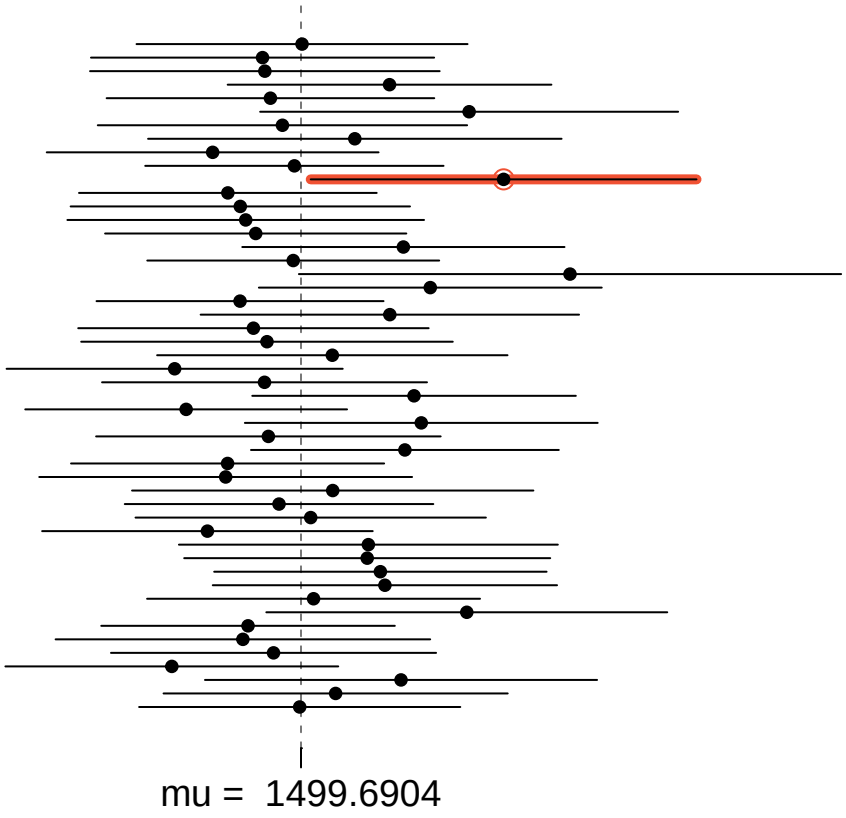
95%, because that is how we defined our confidence interval, i.e. +/- 1.96 standard errors from the sample mean should capture 95% of outcomes.

```
samp_mean <- rep(NA, 50)
samp_sd <- rep(NA, 50)
n <- 60

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 - 1.96 * samp_sd / sqrt(n)
upper_vector <- samp_mean + 1.96 * samp_sd / sqrt(n)

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

```
## [1] 1380.339 1617.128
```

On your own 1: 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))
```



On your own 2: Pick a confidence level of your choosing, provided it is not 95%. What is the appropriate critical value?

Let's choose a confidence interval of 50%, which would mean critical values of +/- 0.67.

```
qnorm(0.25)
```

```
## [1] -0.6744898
```

```
qnorm(0.75)
```

```
## [1] 0.6744898
```

On your own 3: 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?

Using a 50% confidence interval, 27 of the 60 or 45% of the confidence intervals do not include the actual mean. That is pretty close to the 50% we would expect. If we were to increase the number of samples, the percentage should approach 50%.

```
lower_vector <- samp_mean - 0.67 * samp_sd / sqrt(n)
upper_vector <- samp_mean + 0.67 * samp_sd / sqrt(n)

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

```
## [1] 1458.262 1539.205
```

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