STAT 100B Lab 3

Wesley Chang

Summer 2020 Session B

Setup for Lab

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

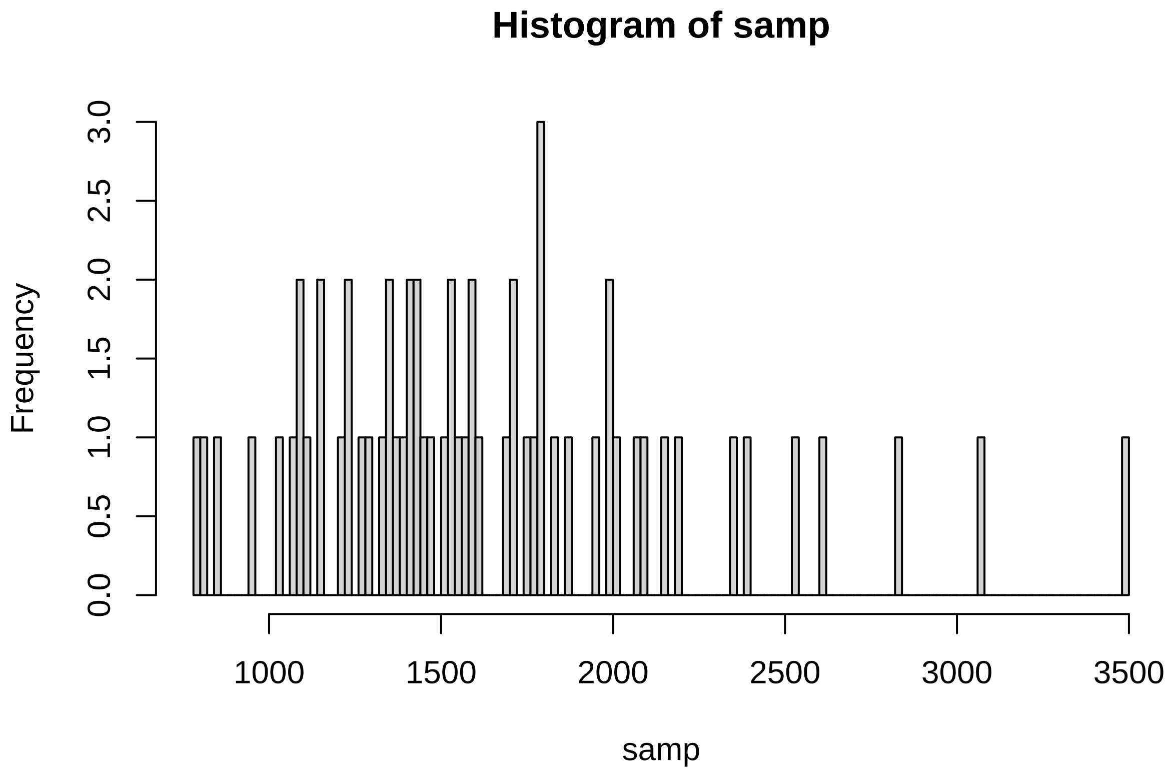
Lab Exercises

Exercise 1

- *Check the distribution of your sample.*
  1. *To build a histogram of your sample, which R code could you use? (2pts)*
  2. *Describe the distribution of your sample. (1pt)*
  3. *What would you say is the typical size within your sample? (1pt)*
  4. *Also state precisely what you interpreted "typical" to mean.(1pt)*
- *Variability from sample to sample*
  1. *Would you expect another student to obtain a distribution identical to yours? (yes/no) (2pts)*
  2. *Woud you expect it to be similar? (yes/no) (2pts)*
  3. *Why or why not? (2pts)*

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

# Generate histogram of sample, with breaks of 100
hist(samp, breaks = 100)
```



```
# summary statistics of samp
summary(samp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     788    1277    1548    1645    1889    3500
```

Answer

The distribution appears to be skewed right, and has a mean of 1644.6833333 and a median of 1548. I would say the typical size within the sample is the median, 1548, as the distribution appears skewed, which would distort the value of the mean. I interpreted typical to be an size that would be represented by the measures of center.

I would expect another student to *not* have a distribution that is identical to mine. However, I would expect it to be similar. Due to sampling variability, it is almost impossible to produce two completely identical sampling distribution, but it is likely they will look similar if the sample size is large enough.

Exercise 2

- *Check the confidence interval you calculated*
  1. *Your confidence interval is? (1pt)*
  2. *Does it capture the true average size of houses in Ames? (yes/no) (1pt)*

```
# calculate mean of the sample
sample_mean <- mean(samp)

# calculate a 95% CI for the sample mean by adding and subtracting the crit values times the standard e
se <- sd(samp)/sqrt(60)
cv <- -qnorm(0.05/2)
lower <- sample_mean - cv*se
upper <- sample_mean + cv*se
c(lower,upper)
```

```
## [1] 1507.222 1782.144
```

```
# true average size of houses in Ames
mean(population)
```

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

Answer

My confidence interval is (1507.2223468, 1782.1443198), which does capture the true average size of houses in Ames, 1499.6904437.

Exercise 3

For the confidence interval to be valid, the sample mean must be normally distributed and have standard error:

$$\frac{s}{\sqrt{n}}$$

What conditions must be met for the normality assumption to be true?

Answer

For the normality assumption to hold, we check for independence (such as a simple random sample) and the success-failure conditions:

$$np \geq 10$$

$$n(1-p) \geq 10$$

Exercise 4

```
pop_mean <- mean(population)
print(pop_mean)
```

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

```
# create empty vectors to save means and standard deviations
samp_mean <- rep(NA,50)
samp_sd <- rep(NA, 50)
n <- 60
```

```
# for loop to calculate means and sd of 50 random samples
num_ci <- 50
for(i in 1:num_ci){
  samp <- sample(population, n)
  samp_mean[i] <- mean(samp)
  samp_sd[i] <- sd(samp)
}

# construct confidence intervals
cv <- -qnorm(0.05/2)
lower_vector <- samp_mean - cv * samp_sd / sqrt(n)
upper_vector <- samp_mean + cv * samp_sd / sqrt(n)

# view first interval
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1371.587 1599.046
```

- *Is the population mean in your first confidence interval?*

  1. *The first interval you obtained is? (1pt)*
  2. *Does your confidence interval capture the population mean? (Yes/No) (1pt)*

Answer

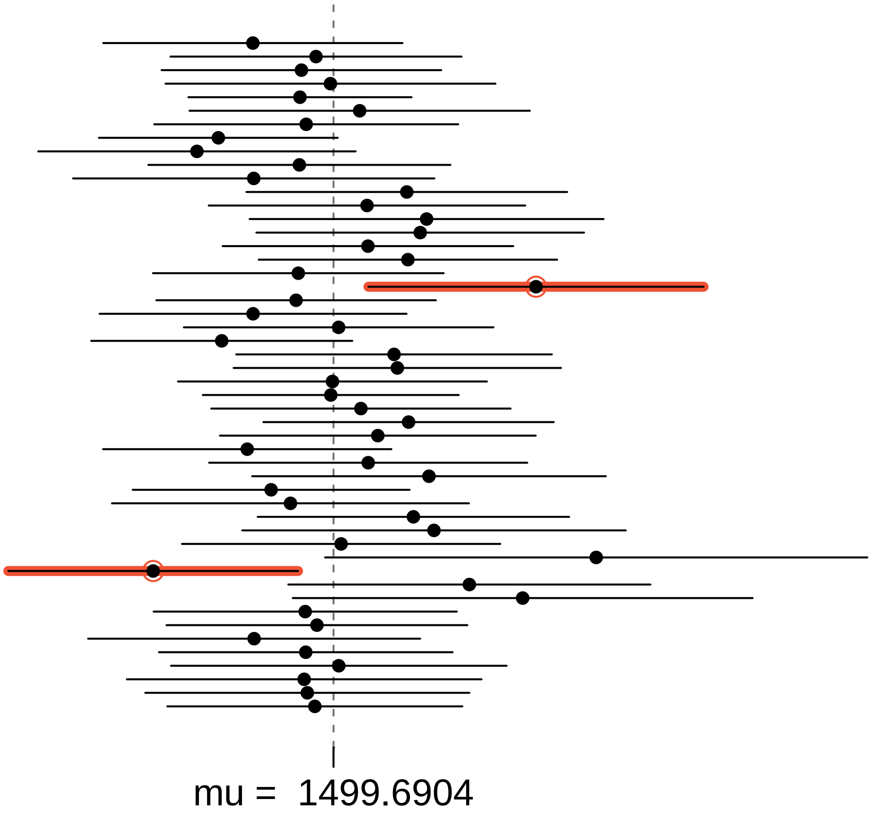
The population mean 1499.6904437 is in my first confidence interval, (1371.5871384,1599.046195). This confidence interval captures the population mean.

Exercise 5

- *What does 95% confidence mean?*

  1. *What proportion of your confidence intervals include the true population mean? (1pt)*
  2. *Is this proportion exactly equal to the confidence level? If not, explain why (2pts)*

```
# plot confidence intervals
plot_ci(lower_vector,upper_vector,mean(population))
```



```
# or, use following code to check what proportion of the confidence intervals capture the true mean
prop_ci <- sum((lower_vector <= pop_mean) & (upper_vector >= pop_mean)) / num_ci
prop_ci
```

```
## [1] 0.96
```

Answer

95% confidence means that there is a 5% chance that my results were generated at random. In other words, if we took many samples and generated confidence intervals for them, about 95% would capture the true population mean. In my sample, 96% of the samples contain the true interval mean, or 48 out of 50 samples. This proportion does not exactly equal the confidence level, as confidence intervals are meant to give a range of values that are likely to have the true parameter value, not be exact.

On your own

Question 1

- *Suppose we'd like to have a 98% confidence interval. What is the appropriate critical value (z value)?*
  1. *Show your R code to find critical value (1pt)*
  2. *The appropriate critical value is? (1pt)*

```r
# 98% critical value for normal distribution
cv2 <- qnorm(.99)

# 98% critical value for normal distribution, method 2
-qnorm(0.02/2)
```

```
## [1] 2.326348
```

Answer

The appropriate critical value for 98% confidence is 2.3263479.

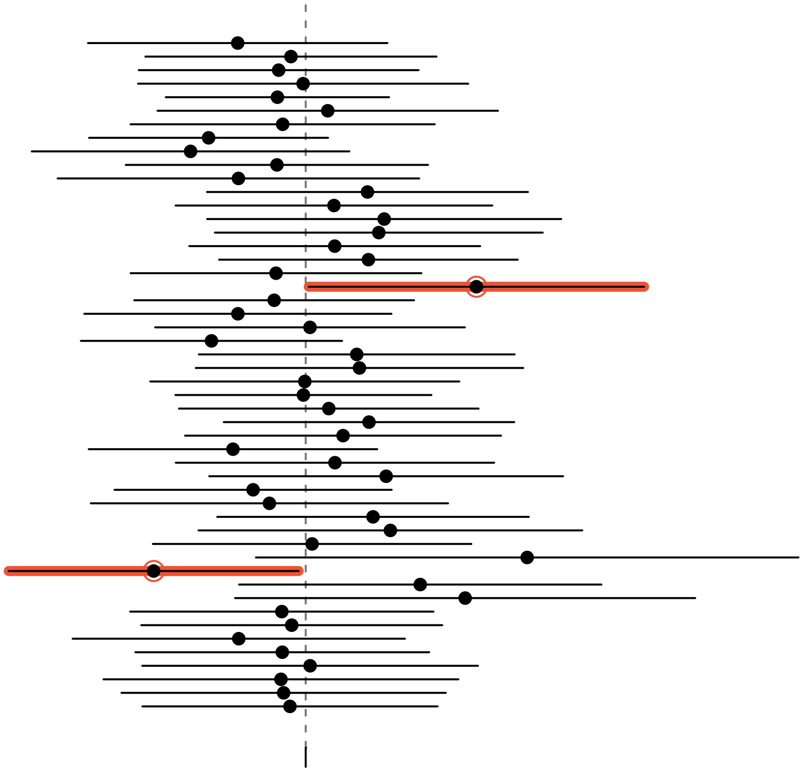
Question 2

- *Construct 50 confidence intervals at the confidence level of 98%. You do not need to obtain new samples, simply calculate new intervals based on the sample means and standard deviations you have already collected.*
  1. *Show your R code to calculate the new intervals (4pts)*
  2. *Using the* `plot_ci` *function, plot all intervales. Show your R code to plot the intervals (2pts)*
  3. *Calculate the proportion of intervals that include the true population mean. Show your R code below (5pts)*

```r
## 1
# construct 50 confidence intervals at level 98%
# since we are using the previously collected samples, do not create new samples
# cv2 defined above as 98% confidence critical value

lower_vector2 <- samp_mean - cv2 * samp_sd / sqrt(n)
upper_vector2 <- samp_mean + cv2 * samp_sd / sqrt(n)

## 2
# plot all intervals, showing R code
plot_ci(lower_vector2, upper_vector2, mean(population))
```



```
## 3
# calculate proportion of intervals that include the true population mean
# since we are using the same samples as above, use num_ci = 50 as the number of samples
prop_ci2 <- sum((lower_vector2 <= pop_mean) & (upper_vector2 >= pop_mean)) / num_ci
```

Answer

The proportion of intervals that include the true population mean at 98% confidence is 96%.