Lab4b

Bin Lin

2016-10-9

```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")
library(IS606)
```

```
##
## Welcome to CUNY IS606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='IS606') will list the demos that are available.
```

```
##
## Attaching package: 'IS606'
```

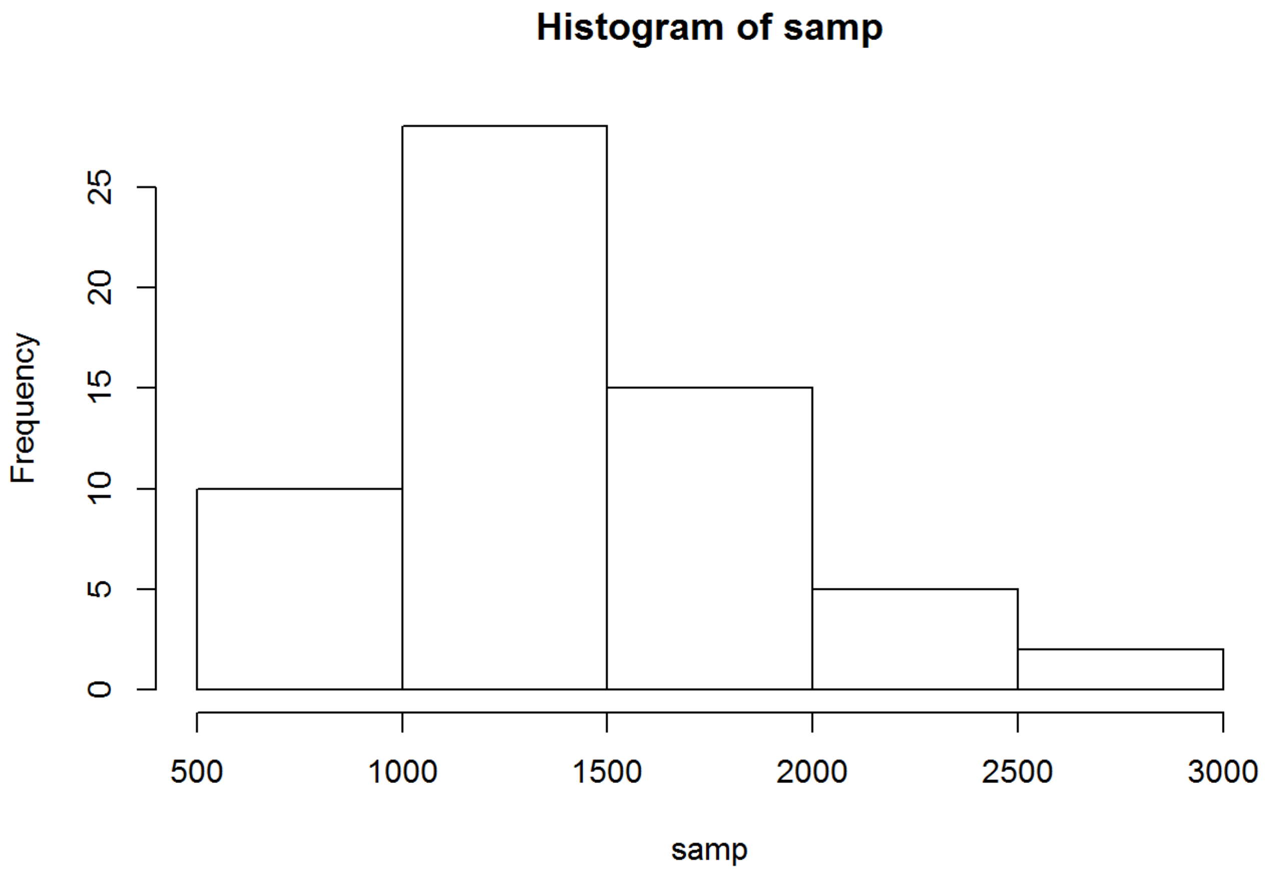
```
## The following object is masked from 'package:utils':
##
##     demo
```

```
load("ames.RData")
set.seed(10)
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     666    1088    1346    1438    1702    2775
```

```
hist(samp)
```



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.

This sample contains elements range from 860 to 3086 It is right skewed with median at 1536, and mean 1635 The median is smaller than the mean, which makes sense because the distribution is right skewed. I think the “typical” size within my sample is the median(1536). Because since the graph is right skewed, only the median will not be influenced by the outliers that are potentially exist in the population. To me, “typical” means the data that can be used to represent the overall situation of the market.

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, I don’t expect the other student will have same exact distribution to mine. But I would expect it to be similar because we are using the same population. The sample size is 60, which means there is quite small standard error of the point estimate. So the distribution of other student will be closed to the mean or median of my sample.

Exercise 3: For the confidence interval to be valid, the sample mean must be normally distributed and have standard error s/n???s/n. What conditions must be met for this to be true? The sample should be independent within groups, and the sample size should be less than 10% of the population size but at least 30. The sample also need to be reasonably symmetric.

Exercise 4: What does “95% confidence” mean? If you’re not sure, see Section 4.2.2. 95% confidence means for all the point estimates (sample mean), 95% of those will fall in the 95% confidence interval (within 2 standard deviation from the population mean). Because each sample mean is likely to be different from each other, therefore, some sample means might be deviated from the true population by large extend. But these sample means only accounted for 5% of the time.

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

```
## [1] 1321.761 1554.806
```

```
mean(population)
```

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

Exercise 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

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%

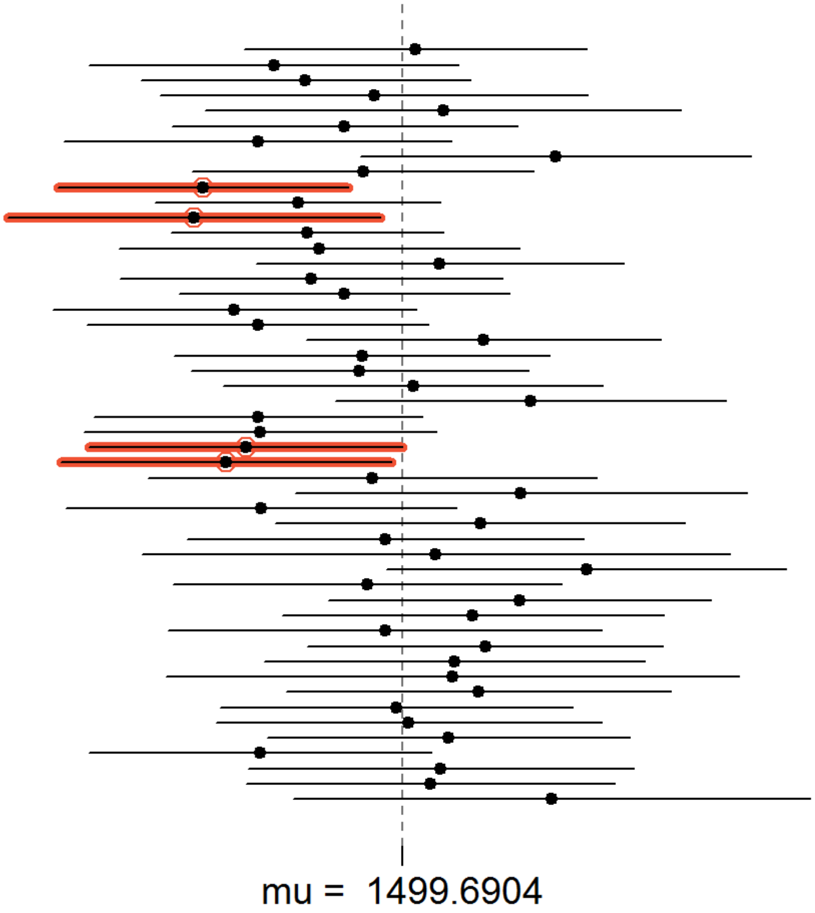
```
samp_mean <- rep(NA, 50)
samp_sd <- rep(NA, 50)
n <- 60
for(i in 1:50){
  samp <- sample(population, n) # obtain a sample of size n = 60 from the population
  samp_mean[i] <- mean(samp)    # save sample mean in ith element of samp_mean
  samp_sd[i] <- sd(samp)        # save sample sd in ith element 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] 1426.654 1771.546
```

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))
```



46/50 = 92% of my confidence intervals include the true population mean. It is not exactly equal to the confidence interval, however it is very closed. Because we only get 50 confidence interval, if we can get infinite condidence interval, the proportion of those that catches population mean will be approaching 95%.

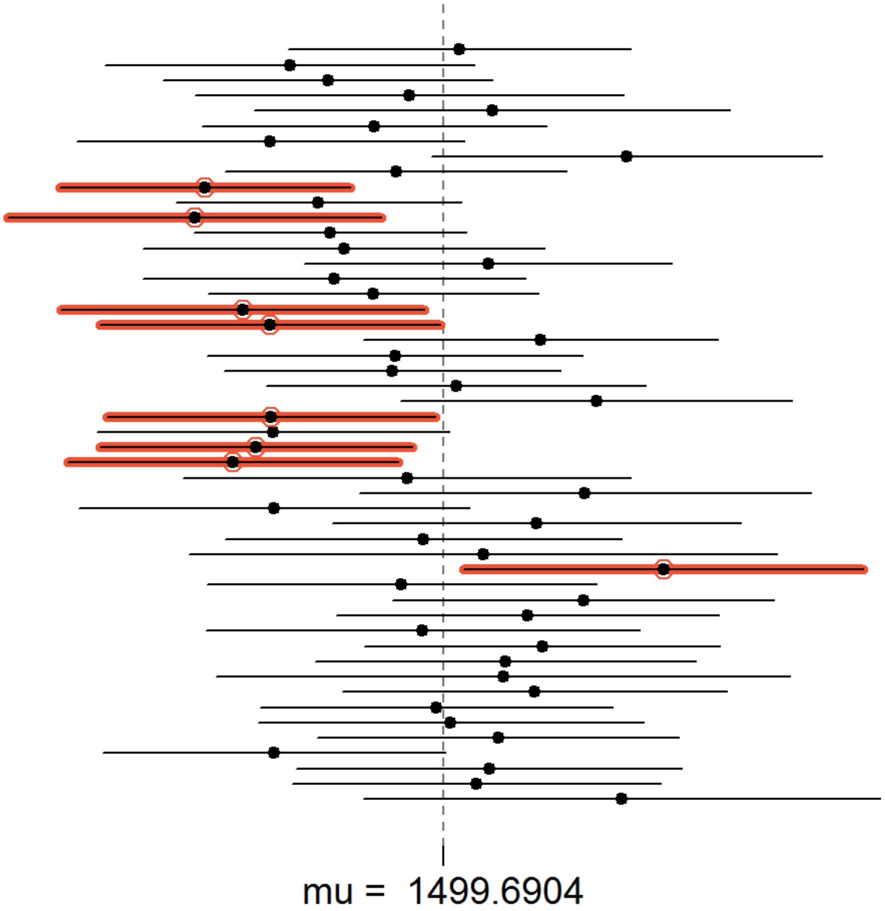
2. Pick a confidence level of your choosing, provided it is not 95%. What is the appropriate critical value? Like 90% confidence interval, the critical value is 1.64

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?

```
lower_vector <- samp_mean - 1.64 * samp_sd / sqrt(n)
upper_vector <- samp_mean + 1.64 * samp_sd / sqrt(n)
c(lower_vector[1], upper_vector[1])
```

```
## [1] 1454.809 1743.391
```

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



From the graph, we know 42/50= 84%, it is still not exatly the same confidence interval that I picked earlier. However, it is still very closed to 90%