Week4-Lab IV

Aamol Gote

9/26/2019

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

load("ames.RData")

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

samp

##  [1]  912  796 1453 1755 1427 1026  816 3194 1319 1822 1792 1718 1396 2032
## [15] 1848 1920 1428  672 1616 1852  980  968 1850 1694 2048 1422 1632  936
## [29] 1528 1560 1302 1710 1118 1797 1245 1141 1297 1460 1948 1365 1572  784
## [43] 1604 1456 1337 1595 1033 2202 1870 2515  960 1144 2452 1284  924  816
## [57] 1162  894 1445 1113
```

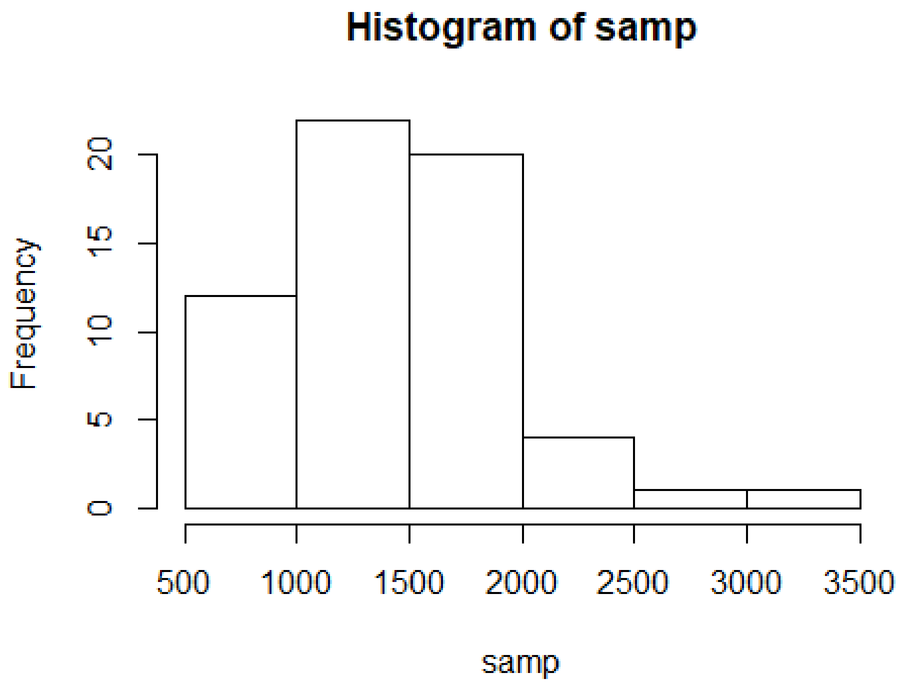
Excercise 1

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(samp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     672    1117    1436    1466    1764    3194

hist(samp)
```



- Population data of houses from Ames is unimodal and right skewed.
- Range is from 672 to 3194. Median is 1436 and Mean is 1466
- For the sample size of 60 (n=60), typical size resembles the sample mean which in this case is 1466. So sample mean 1466 is an average value that represents the average size of the house in Ames in Iowa.

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?

Another student's distribution would **not be exactly identical**, but it would be somewhat similar. Reason been this is **randomly sampled** and another student may have different set of samples but as samples are taken from the same set of population there would be similarities between different samples.

Exercise 3

For the confidence interval to be valid, the sample mean must be normally distributed and have standard error s/√n what conditions must be met for this to be true?

1. Independent observations
2. Sampling size greater than 30

3. Data Not strongly skewed

All above are part of CLT (Central limit theorem) informal description

Exercise 4

What does “95% confidence” mean? If you’re not sure, see Section 4.2.2.

Suppose we take many samples and calculate confidence interval for each sample then 95% of the time the mean would be in between the calculated confidence interval for each sample.

```
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?

point estimate +- 1:96 * SE

SE = SD/√n Sample Size = n = 60

Below is the 95% confidence range

```
sample_mean <- mean(samp)
se <- sd(population)/(sqrt(60))
lowerbound <- sample_mean - (1.96 * se)
upperbound <- sample_mean + (1.96 * se)
c(lowerbound, upperbound)

## [1] 1338.039 1593.861
```

Actual mean

```
mean(population)

## [1] 1499.69
```

Actual mean falls in between the confidence interval range and captures the true average size of houses.

Neighbor’s confidence interval should also capture the mean value in between them.

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% confidence interval should capture the true population mean.
- For each student the sample is going to be different so confidence interval are going to vary but for the 95% of the students should capture the true population mean. Suppose class contains 100 students take the sample size of 60 and calculate the confidence interval, then 95 students should capture the true population mean.
- We are considering the confidence interval for 95% for above calculation based on which we are taking the value 1.96 which is the Z-value for 95% confidence interval.

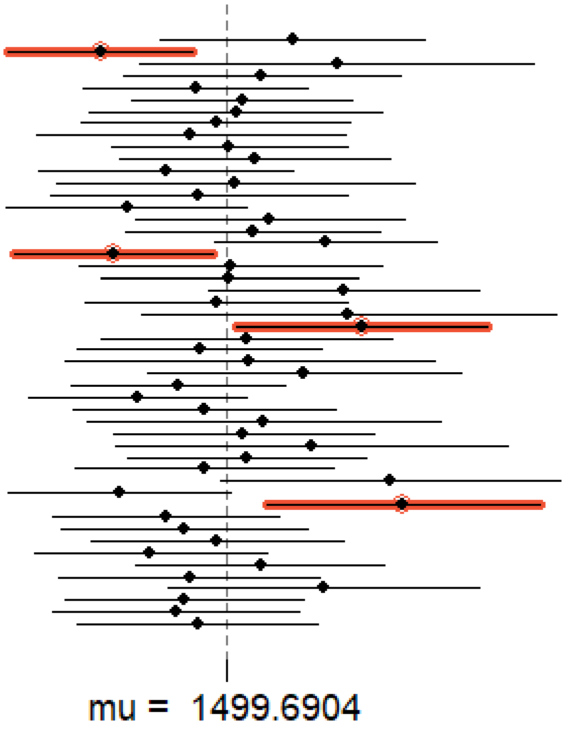
```
samp_mean <- rep(NA, 50)
samp_sd <- rep(NA, 50)
n <- 60
for(i in 1:50){
  samp <- sample(population, n) # obtain a sample of size n = 60 from the
population
  samp_mean[i] <- mean(samp)    # save sample mean in ith element of
samp_mean
  samp_sd[i] <- sd(samp)        # save sample sd in ith element 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] 1357.542 1585.558
```

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))
```



- 50 samples of sample size 60
- 4 out of 50 contain the true mean, which percentage wise is 1 – 4/50 = 0.92 %.
- So the confidence level is 92%, which is not exactly to the selected confidence interval of 95%. Confidence interval are never exact match and are meant to be approximation.

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

   Confidence level = 90%

```
confidence_level <- 90
signifiance_level  <- 1-(confidence_level/100)
cp <- 1-(signifiance_level/2)
qnorm(cp)

## [1] 1.644854
```

   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

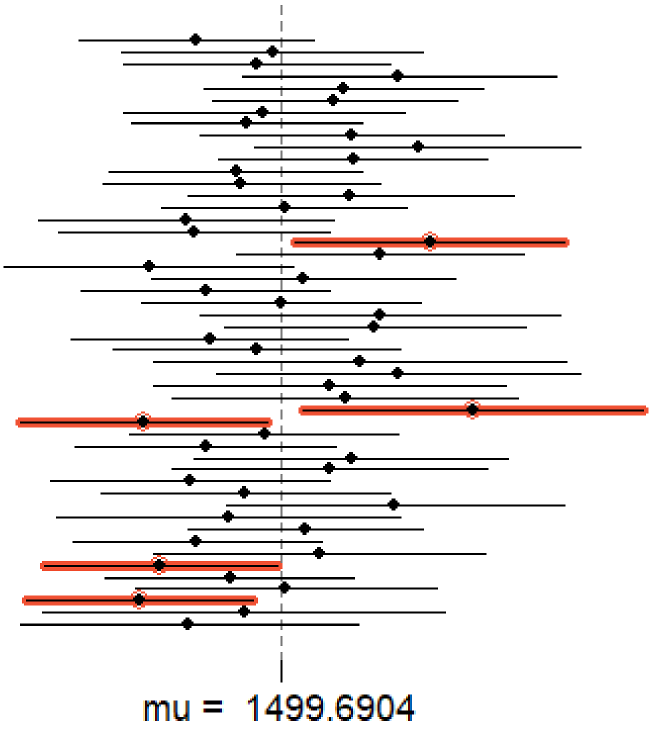
plot_ci function, plot all intervals and calculate the proportion of intervals that include the true population mean. How does this percentage compare to the confidence level selected for the intervals?

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

```
## [1] 1310.018 1554.049
```

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



- Proportion of Confidence interval plot at 90% confidence interval is 1 – 5/50 = 0.90 = 90% which is exact match to 90% selected confidence level, but anything approximately near number to 90% confidence level should be fine as well.