Lab4

Priya Shaji

3/2/2019

## A) Foundations for statistical inference - Sampling distributions

### Lab

#### The data

Download the file and Load the data

#library(DATA606)  
#download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")  
load("/Users/priyashaji/Documents/cuny msds/Spring'19/data 606/Labs/Lab4/ames.RData")

Renaming columns to be used

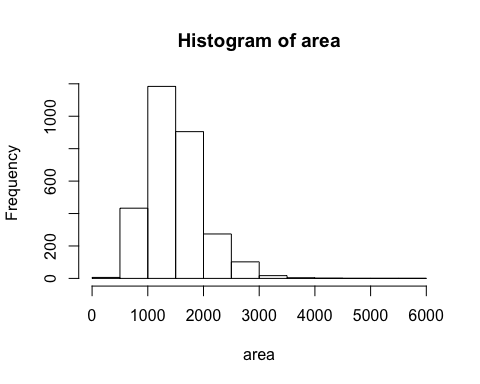
area <- ames$Gr.Liv.Area  
price <- ames$SalePrice

Let’s look at the distribution of area in our population of home sales by calculating a few summary statistics and making a histogram.

summary(area)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 1126 1442 1500 1743 5642

hist(area)



mean(area)

## [1] 1499.69

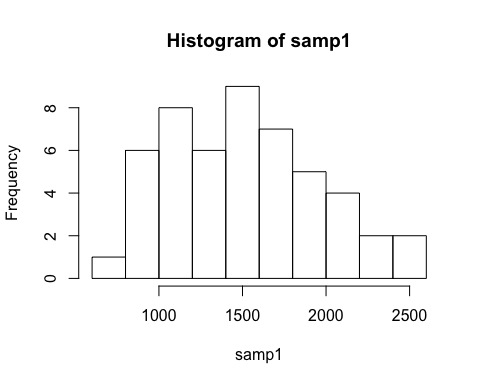
#### The unknown sampling distribution

If we were interested in estimating the mean living area in Ames based on a sample, we can use the following command to survey the population.

samp1 <- sample(area, 50)  
summary(samp1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 729 1138 1460 1510 1796 2490

hist(samp1)



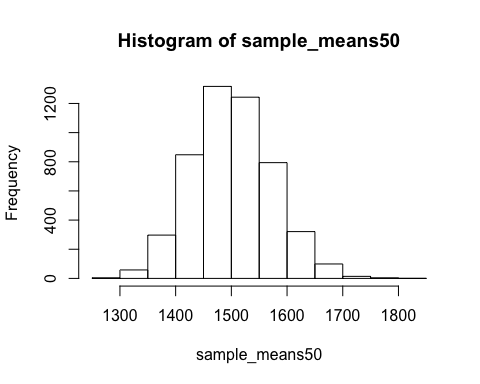
If we’re interested in estimating the average living area in homes in Ames using the sample, our best single guess is the sample mean.

mean(samp1)

## [1] 1510.1

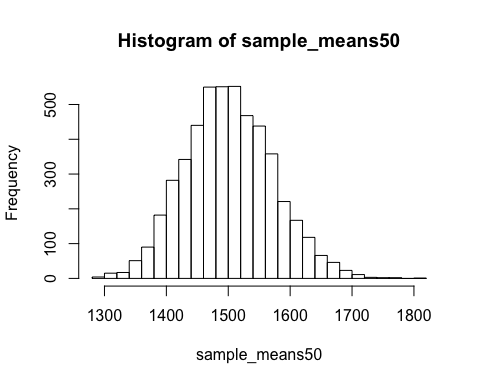
Generate 5000 samples and compute the sample mean of each.

sample\_means50 <- rep(NA, 5000)  
  
for(i in 1:5000){  
 samp <- sample(area, 50)  
 sample\_means50[i] <- mean(samp)  
 }  
  
hist(sample\_means50)



If you would like to adjust the bin width of your histogram to show a little more detail, you can do so by changing the breaks argument.

hist(sample\_means50, breaks = 25)



#### Interlude: The for loop

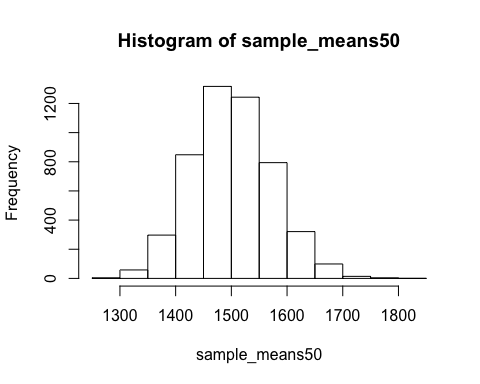
Calculating the sample mean without the for loop

With the for loop, these thousands of lines of code are compressed into a handful of lines. We’ve added one extra line to the code below, which prints the variable i during each iteration of the for loop. Run this code.

#### Sample size and the sampling distribution

Mechanics aside, let’s return to the reason we used a for loop: to compute a sampling distribution, specifically, this one.

hist(sample\_means50)

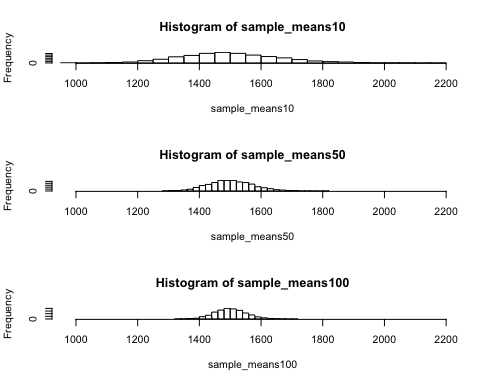


To get a sense of the effect that sample size has on our distribution, let’s build up two more sampling distributions: one based on a sample size of 10 and another based on a sample size of 100.

sample\_means10 <- rep(NA, 5000)  
sample\_means100 <- rep(NA, 5000)  
  
for(i in 1:5000){  
 samp <- sample(area, 10)  
 sample\_means10[i] <- mean(samp)  
 samp <- sample(area, 100)  
 sample\_means100[i] <- mean(samp)  
}

To see the effect that different sample sizes have on the sampling distribution, plot the three distributions on top of one another.

par(mfrow = c(3, 1))  
  
xlimits <- range(sample\_means10)  
  
hist(sample\_means10, breaks = 20, xlim = xlimits)  
hist(sample\_means50, breaks = 20, xlim = xlimits)  
hist(sample\_means100, breaks = 20, xlim = xlimits)



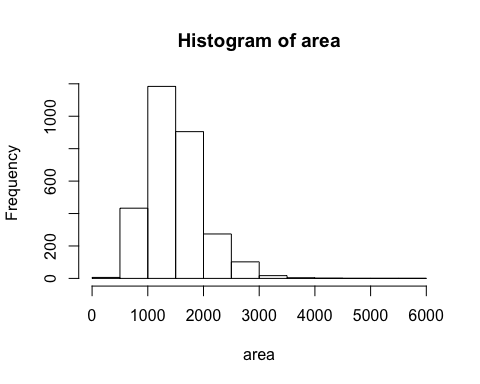
### Exercises

#### Exercise 1

Describe this population distribution.

Answer 1)

hist(area)



The distribution is left skewed and unimodal

#### Exercise 2

Describe the distribution of this sample. How does it compare to the distribution of the population?

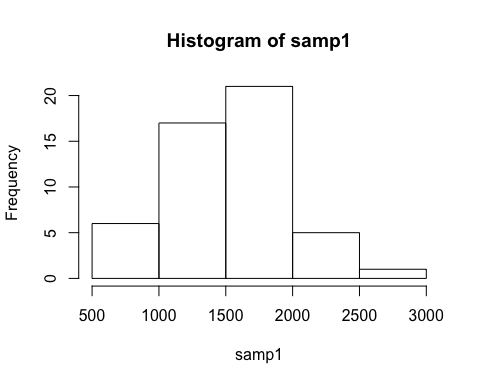
samp1 <- sample(area, 50)  
summary(samp1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 725 1230 1574 1545 1726 2956

mean(samp1)

## [1] 1544.86

hist(samp1)



The distribution of the sample seems to be unimoadl and left skewed. It looks similar to the distribution of the population.

#### Exercise 3

Take a second sample, also of size 50, and call it samp2. How does the mean of samp2 compare with the mean of samp1? Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population mean?

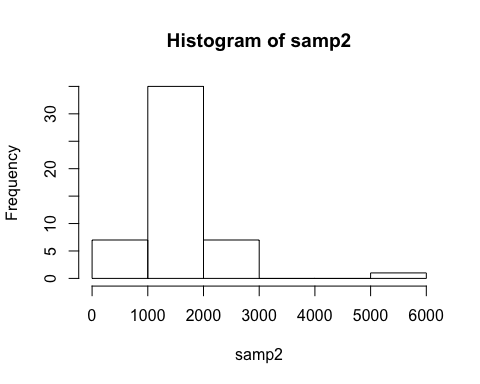
samp2 <- sample(area, 50)  
summary(samp2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 540 1139 1378 1547 1764 5642

mean(samp2)

## [1] 1547

hist(samp2)



Mean of sample 1: 1537.16

Mean of sample 2: 1616.56

Sample size : 100

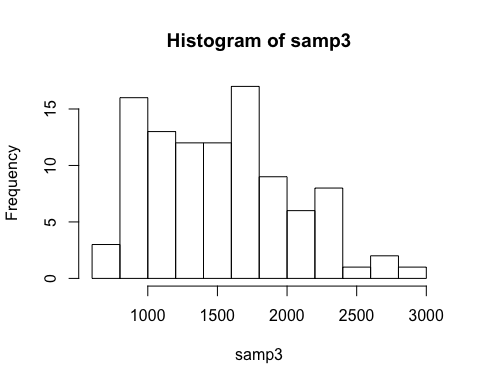
samp3 <- sample(area, 100)  
summary(samp3)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 672 1088 1446 1523 1872 2822

mean(samp3)

## [1] 1522.87

hist(samp3)



Sample size: 1000

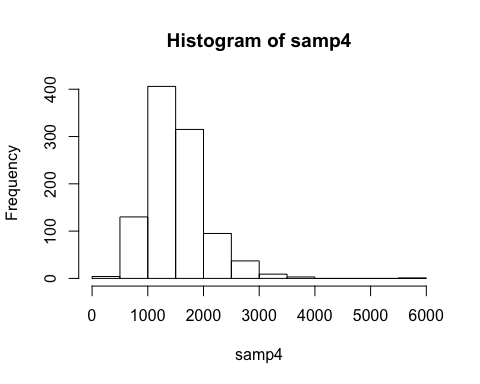
samp4 <- sample(area, 1000)  
summary(samp4)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 1144 1456 1517 1768 5642

mean(samp4)

## [1] 1517.246

hist(samp4)



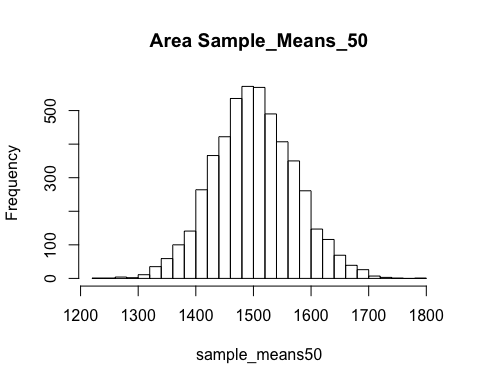
The mean of the samp2 will not be the same as the mean of samp1. It is difficult to determine if samp2 will consistantly have a higher or lower mean because the samples are randomly selected.

The larger the sample mean, the more accurate the estimate of the population mean will be. As such, the sample size of 1,000 will be the closest estimate to the original population.

#### Exercise 4

How many elements are there in sample\_means50? Describe the sampling distribution, and be sure to specifically note its center. Would you expect the distribution to change if we instead collected 50,000 sample means?

sample\_means50 = rep(NA, 5000)  
  
for(i in 1:5000){  
 samp = sample(area, 50)  
 sample\_means50[i] = mean(samp)  
}  
  
hist(sample\_means50, breaks = 25, main = "Area Sample\_Means\_50")



The sampling distribution is normal and unimodal with a center of approximately 1500. If sample means is 50,000, the more accurate the estimate of the population mean will be. And sample size of 1,000 will be the closest estimate to the original population.

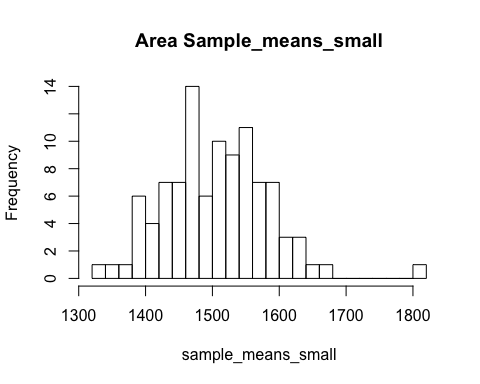
#### Exercise 5

To make sure you understand what you’ve done in this loop, try running a smaller version. Initialize a vector of 100 zeros called sample\_means\_small. Run a loop that takes a sample of size 50 from area and stores the sample mean in sample\_means\_small, but only iterate from 1 to 100. Print the output to your screen (type sample\_means\_small into the console and press enter). How many elements are there in this object called sample\_means\_small? What does each element represent?

sample\_means\_small = rep(NA, 100)  
  
for(i in 1:100){  
 samp = sample(area, 50)  
 sample\_means\_small[i] = mean(samp)  
}  
sample\_means\_small

## [1] 1412.02 1464.52 1581.26 1386.78 1456.38 1566.36 1500.42 1547.04  
## [9] 1473.58 1513.84 1550.32 1605.60 1418.70 1640.60 1590.20 1543.84  
## [17] 1360.68 1572.30 1520.84 1504.86 1579.36 1540.20 1439.78 1466.94  
## [25] 1814.02 1453.62 1520.38 1503.04 1434.72 1472.68 1582.92 1442.66  
## [33] 1339.42 1536.80 1497.38 1414.34 1623.14 1583.52 1545.92 1417.04  
## [41] 1521.08 1459.22 1503.02 1539.94 1475.26 1512.46 1546.80 1430.44  
## [49] 1584.80 1581.98 1571.72 1551.96 1571.88 1513.84 1381.96 1540.70  
## [57] 1475.52 1626.14 1591.26 1428.40 1388.60 1446.94 1483.84 1458.66  
## [65] 1546.30 1476.42 1471.62 1618.48 1502.44 1354.92 1625.78 1447.82  
## [73] 1384.00 1492.74 1471.98 1667.48 1472.46 1529.98 1436.08 1561.64  
## [81] 1495.18 1569.66 1539.54 1546.60 1548.62 1532.30 1501.20 1462.36  
## [89] 1422.62 1501.56 1539.00 1491.46 1490.28 1474.72 1391.82 1423.62  
## [97] 1610.68 1462.00 1386.00 1471.68

hist(sample\_means\_small, breaks = 25, main = "Area Sample\_means\_small")



There are 100 elements in sample\_means\_small object.

Each element represents the mean of a sample of size 50, drawn for the population area.

#### Exercise 6

When the sample size is larger, what happens to the center? What about the spread?

When the sample size is larger, the spread decreases and the center grows closer to the population area’s center.

### On your Own

#### Ques\_1

Take a random sample of size 50 from price. Using this sample, what is your best point estimate of the population mean?

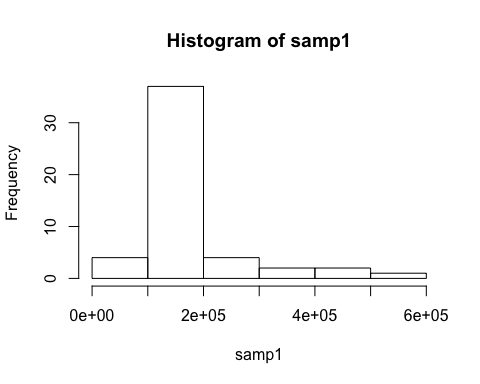
price <- ames$SalePrice  
samp1 <- sample(price, 50)  
summary(samp1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 79000 123000 143250 175735 179500 591587

mean(samp1)

## [1] 175734.9

hist(samp1)



The mean is the best point estimate for the population mean.

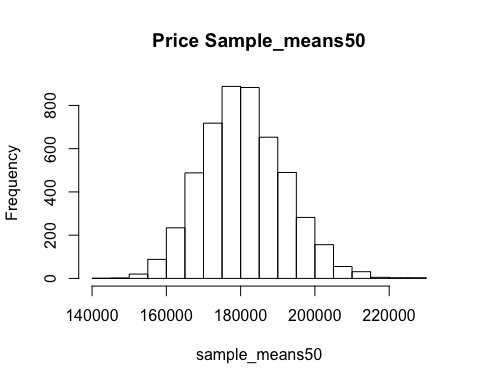
#### Ques\_2

Since you have access to the population, simulate the sampling distribution for x¯price by taking 5000 samples from the population of size 50 and computing 5000 sample means. Store these means in a vector called sample\_means50. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean home price of the population to be? Finally, calculate and report the population mean.

sample\_means50 = rep(NA, 5000)  
  
for(i in 1:5000){  
 samp = sample(price, 50)  
 sample\_means50[i] = mean(samp)  
}  
summary(sample\_means50)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 144427 173026 180241 180841 188127 226565

hist(sample\_means50, breaks = 25, main = "Price Sample\_means50")



Shape of the sampling distribution is normal and unimodal. The mean home price of the population seems to be 181121.

Calculating mean for population:

summary(price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12789 129500 160000 180796 213500 755000

The actual mean of the price is 180796. My estimate was higher than the actual mean but much closer to the actual value.

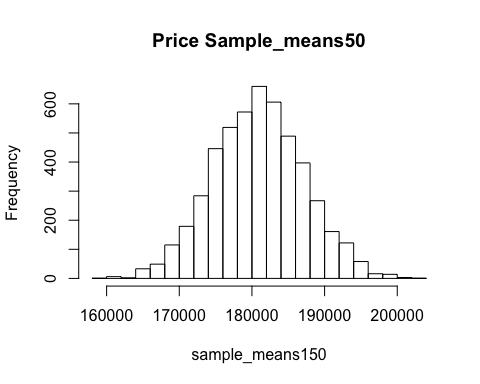
#### Ques\_3

Change your sample size from 50 to 150. Repeat process above but store these means in a vector called “sample\_means150”. Describe the shape of this sampling distribution. Compare it to the sampling distribution for a sample size of 50. Based on this sampling distribution, what would you guess to be the mean sale price of homes in Ames?

sample\_means150 = rep(NA, 5000)  
  
for(i in 1:5000){  
 samp = sample(price, 150)  
 sample\_means150[i] = mean(samp)  
}  
summary(sample\_means150)

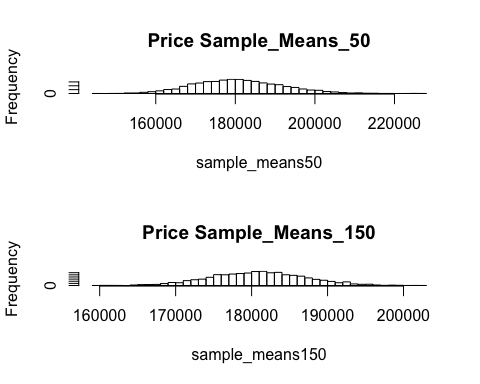
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 159397 176470 180909 180886 185085 202223

hist(sample\_means150, breaks = 25, main = "Price Sample\_means50")



The sample distribution is normal and unimodal.

# Comparisons between means50 and means150  
par(mfrow= c(2,1))  
hist(sample\_means50, breaks = 50, main = "Price Sample\_Means\_50")  
hist(sample\_means150, breaks = 50, main = "Price Sample\_Means\_150")



summary(sample\_means50)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 144427 173026 180241 180841 188127 226565

summary(sample\_means150)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 159397 176470 180909 180886 185085 202223

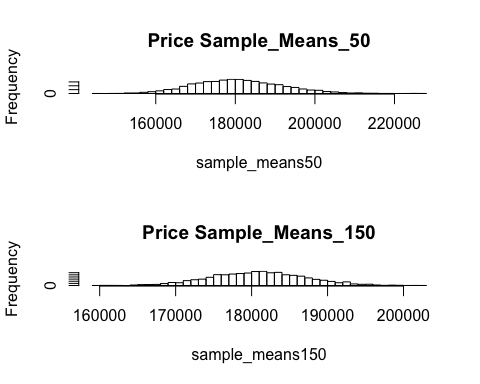
The sampling distribution for sample size 50 is more spread as compared to the sample size of 150.

The mean sale price seems to be 180725, which is approx the actual mean sale price of homes in ames.

#### Ques\_4

Of the sampling distributions from 2 and 3, which has a smaller spread? If we’re concerned with making estimates that are more often close to the true value, would we prefer a distribution with a large or small spread?

par(mfrow= c(2,1))  
hist(sample\_means50, breaks = 50, main = "Price Sample\_Means\_50")  
hist(sample\_means150, breaks = 50, main = "Price Sample\_Means\_150")



Of the two sampling distributions (sample\_means50 and sample\_means150), the second (sample\_means150) has the smaller spread.

It is preferable to have a distribution with a smaller spread when trying to make estimates that are close to the true value, because there is less uncertainty of where the true estimates lie.

## B) Foundations for statistical inference - Confidence intervals

### Lab

#### The data

Download the file and Load the data

#library(DATA606)  
#download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")  
load("/Users/priyashaji/Documents/cuny msds/Spring'19/data 606/Labs/Lab4/ames.RData")

Let’s start with a random sample size of 60

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

#### 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)

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] 1266.355 1505.779

#### Confidence levels

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

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.

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.

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.

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.

c(lower\_vector[1], upper\_vector[1])

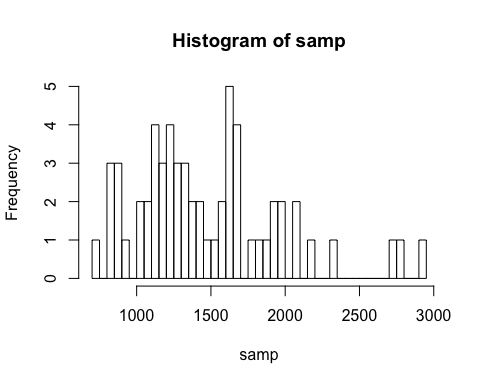
## [1] 1244.038 1472.162

### Exercises

#### 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.

hist(samp,breaks = 50)



The population unimodal and left skewed.

The typical size within a sample ranges within 1300 to 1450

By “typical size” I analyzed the size which has highest number of frequency.

#### 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 will bw different,as the sample of population might vary for different students and also it is relatively small sample size.

The sample size is 60 out of 5000 cases.The distributions can be similar on most rare occasions.

#### 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?

The following condtions should be met:

1. The observations should be independent.
2. The sample size should be more than 30.
3. The sample data should not be strongly skewed.

#### Exercise 4

What does “95% confidence” mean?

It means that we are 95 % sure about the true population mean lying between interval (-2 Sd - +2 Sd) when ploting sampling distribution. See the range below:

lower

## [1] 1266.355

upper

## [1] 1505.779

#### 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?

summary(population)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 1126 1442 1500 1743 5642

Yes, if you see the mean of the population above, it does fall under lower and upper level of the sample calculated using 95% confidence.

#### 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.

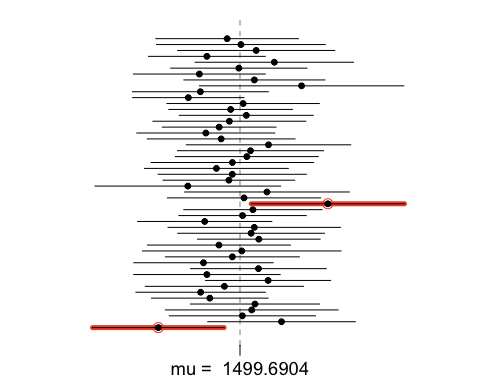
95% of the students will capture true population because of the fact that everyone of them would have used the following (Sample Mean + (+/- 1.96 \* SE))

### On your own

#### Ques\_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))



In the first run I got 46/50 i.e. 92% which is lower than 95%. The 95% confidence level does not mean that every run would have poopulation mean lying between the sample mean interval. There would be cases where the population mean would be lying outside interval (Sample Mean + (+/- 1.96 \* SE)).

#### Ques\_2

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

We can choose the confidence level of 99% by following the given formula: (Sample Mean + (+/- 2.58 \* SE)).

#### Ques\_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 - 2.58 \* samp\_sd / sqrt(n)   
upper\_vector <- samp\_mean + 2.58 \* samp\_sd / sqrt(n)  
plot\_ci(lower\_vector, upper\_vector, mean(population))

