Lab5

James Law

April 3, 2018

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

### Exercise 1:Describe the shape, center (mean), and spread (standard deviation) of this population distribution.

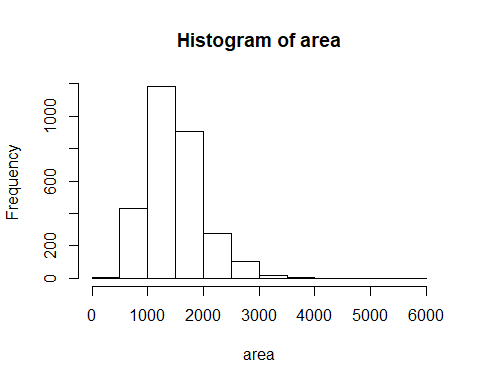
The shape of this distribution is unimodal and slightly right skewed, with the center measured by the mean at 1500 square feet and the spread measured by the standard deviation at 505.5 square feet.

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

summary(area)

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

hist(area)



sd(area)

## [1] 505.5089

### Exercise 2:Calculate summary statistics and plot a histogram of your sample. Describe the shape, center (mean), and spread (standard deviation) of this sample distribution. How does it compare to the population distribution you described in Exercise 1?

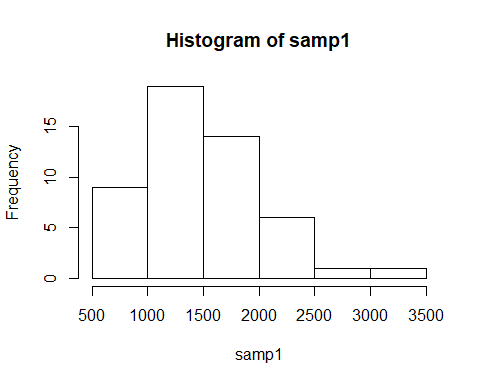
The shape of this sample distribution is right skewed just like the population distribution. The center for the sample disribtion is 1471 square feet, pretty close compared to 1500 square feet for the population distribuition. Lastly the spread measured by the standard dievation for the sample is 510 square feet, still close to the population spread at 505.5 square feet. Based on these measurements for center and spread I would say our sample resembles the population.

set.seed(99)  
samp1 <- sample(area, 50)

summary(samp1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 764 1063 1405 1471 1735 3005

hist(samp1)



sd(samp1)

## [1] 510.992

### Exercise 3:Take a second sample, also of size 50, and name 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? Why?

The second sample mean is very close to the mean of the population distribution. If we had two sample sizes of 100 and 1000 the sample with size 1000 would be more accurate as the size is closer to the actual population size.

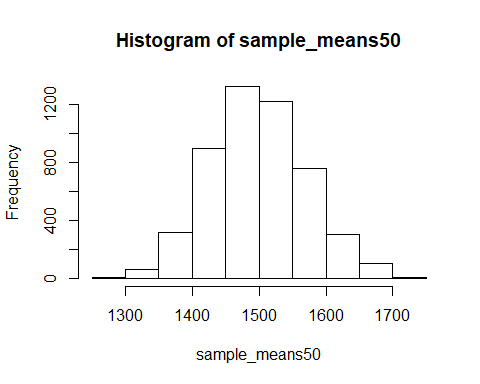
set.seed(98)  
samp2 <- sample(area, 50)  
  
mean(samp1)

## [1] 1471.26

mean(samp2)

## [1] 1503.66

set.seed(100)  
sample\_means50 <- rep(0, 5000)   
for (i in 1:5000) {   
 samp <- sample(area, 50)   
 sample\_means50[i] <- mean(samp)   
}  
hist(sample\_means50)



### Exercise 4:How many elements are there in sample\_means50? Describe the shape, center (mean), and spread (standard deviation) of the sampling distribution. How would you expect the sampling distribution to change if we instead collected 50,000 sample means?

There are 5000 elements in sample\_means50 because we looped from 1 to 5000. The shape of the sampling distribution is aproximitely normal, with the center measured by the mean at 1499 square feet and the spread measured by the standard dievation at 71.2 square feet. If we collected 50,000 sample means instead of 5000, I believe the graph would be more normal with a very slightly tighter spread, but the mean would stay where it is at.

summary(sample\_means50)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1274 1448 1496 1499 1547 1749

sd(sample\_means50)

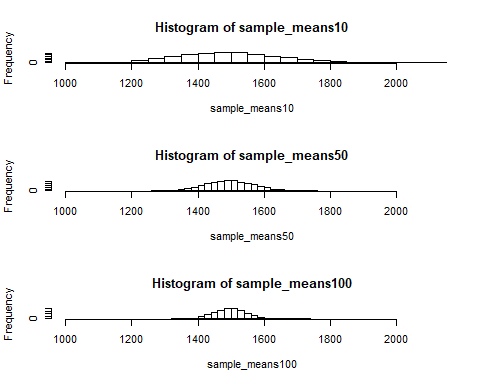
## [1] 71.1843

### Exercise 5: When the sample size is larger, what happens to the center (mean) of the sampling distribution? What about the spread (standard deviation)?

As sample size gets the larger the center measured by the mean stays the same, however the spread measured by the standard dievation gets tighter from 157 square feet for sample\_means10 to 50.2 for sample\_means100.

set.seed(101)  
sample\_means10 <- rep(0, 5000)   
sample\_means100 <- rep(0, 5000)   
for (i in 1:5000) {   
 samp <- sample(area, 10)   
 sample\_means10[i] <- mean(samp)   
 samp <- sample(area, 100)   
 sample\_means100[i] <- mean(samp)   
}

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)



sd(sample\_means10)

## [1] 157.2376

sd(sample\_means50)

## [1] 71.1843

sd(sample\_means100)

## [1] 50.16119

### Homework Assignmnets

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

set.seed(100)  
price\_samp <- sample(price, 50)  
mean(price\_samp)

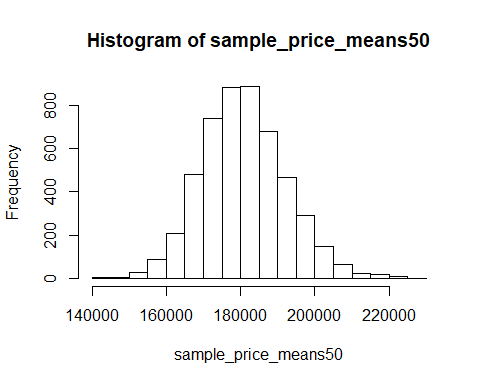
## [1] 174131.4

$174131.4 is our best point estimate of the population mean for home price.

### 2. Since you have access to the population, simulate the sampling distribution for 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.

Based on this simulated sampling distribution I would guess the mean home price for the population would be $180,905, where the actual mean is 180,796.10

set.seed(101)  
sample\_price\_means50 <- rep(0, 5000)   
for (i in 1:5000) {   
 samp <- sample(price, 50)   
 sample\_price\_means50[i] <- mean(samp)   
}  
hist(sample\_price\_means50)



mean(sample\_price\_means50)

## [1] 180905

sd(sample\_price\_means50)

## [1] 11246.9

mean(price)

## [1] 180796.1

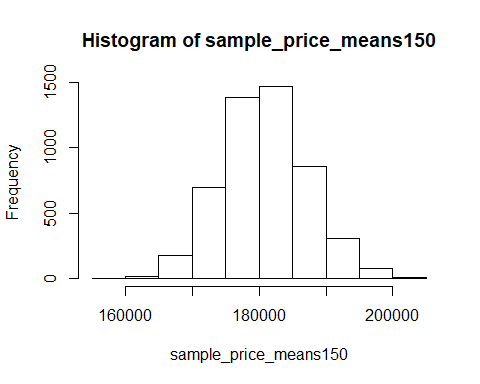
mean(price)

## [1] 180796.1

### 3. Change your sample size from 50 to 150, then compute the sampling distribution using the same method as above, and store these means in a new vector called sample\_means150. Describe the shape of this sampling distribution, and 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?

Based on my new sampling distribution of n=150 I would guess that the mean sale price in Ames would be $180,844.30, close to both the population mean of 180,796.10 and my sampling disribution where n=50 of 180,905.

set.seed(102)  
sample\_price\_means150 <- rep(0, 5000)   
for (i in 1:5000) {   
 samp <- sample(price, 150)   
 sample\_price\_means150[i] <- mean(samp)   
}  
hist(sample\_price\_means150)



mean(sample\_price\_means150)

## [1] 180844.3

sd(sample\_price\_means150)

## [1] 6298.131

### 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?

The sampling distribution where n=150 has the tighter spread with the standard dievation of 6,298.1, compared to the spread of the sampling distribution where n=50 where the standard dievation is 11,246.9. If we’re concerned with making estimates that are more often close to the true value, we would prefer a sampling distribution with a small spread, because there is less variation in the sample means meansing our sample mean would be closer to the population mean.