Lab5

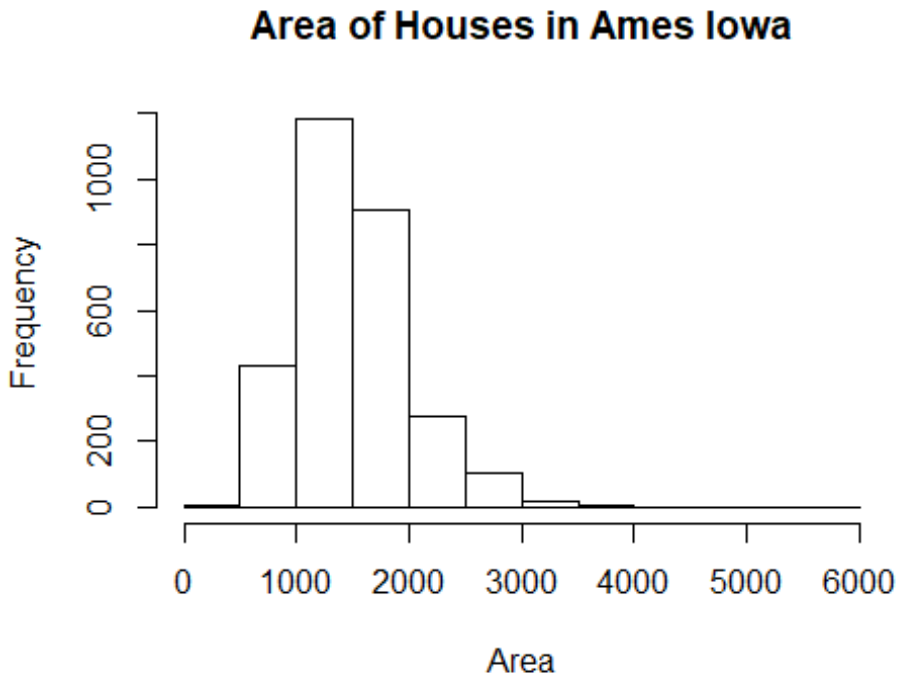
```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "am
es.RData")
load("ames.RData")

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

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

```
summary(area)
```

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

```
hist(area, main = "Area of Houses in Ames Iowa", xlab = "Area")
```




```
mean(area) #The mean of the area
```

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

```
sd(area) #The standard deviation of the area
```

```
## [1] 505.5089
```

The shape of the data is right skewed. The mean is 1499.69, while the standard deviation is 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?

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     630    1187    1409    1448    1654    3086
```

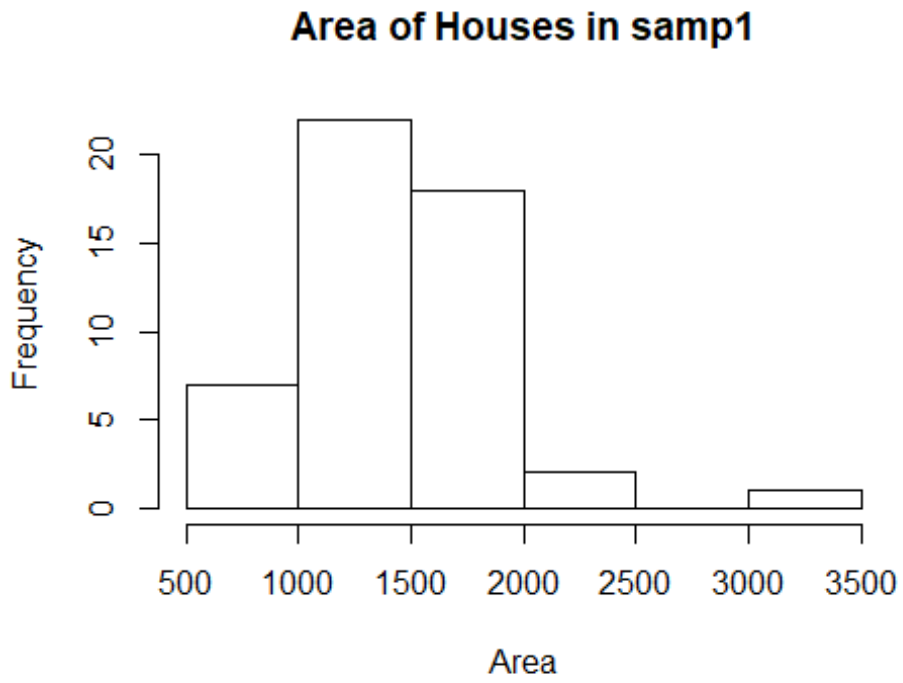
```
mean(samp1) #The mean of samp1
```

```
## [1] 1447.52
```

```
sd(samp1) #The standard deviation of the area
```

```
## [1] 429.4866
```

```
hist(samp1, main = "Area of Houses in samp1", xlab = "Area")
```



The shape of the data skewed right. The mean is 1447.52 while the standard deviation is 429.4866. While comparing this population to Exercise 1, The mean and the standard deviation are smaller with Example 2 and the population is skewed right.

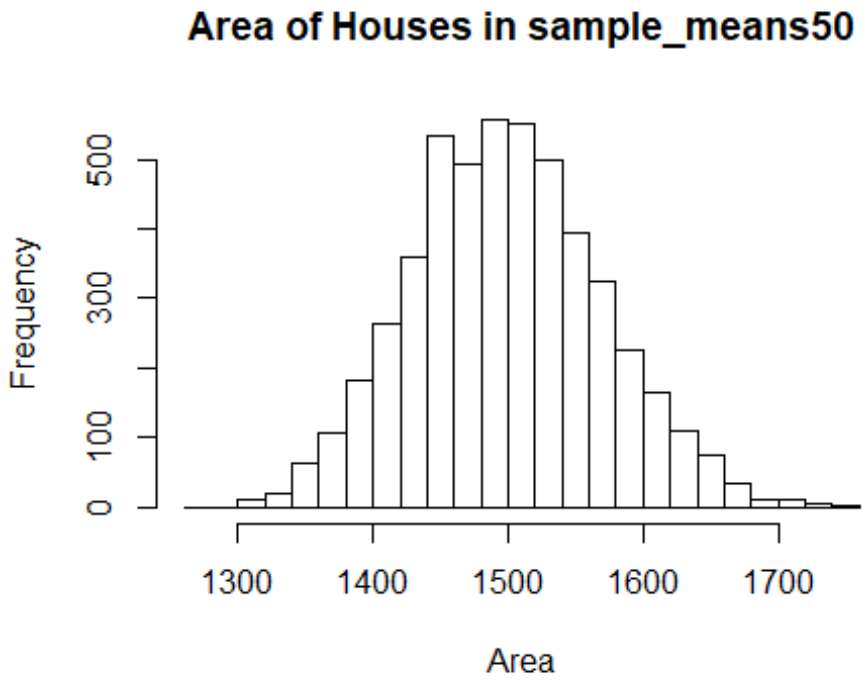
Exercise 3: Take a second sample, also of size 50, and name it samp2. How does the mean of samp2 compare with the mean of samp1? Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population mean? Why?

```
samp2 <- sample(area, 50)
mean(samp2) #The mean of samp2

## [1] 1613.2
```

The mean of sample 2 is 1613.2. The mean is larger than sample 1. If we took 2 more samples of size 100 and 1000, the size of 1000 would provide a more accurate estimate of the population mean. With a larger sample size, its more likely we will have enough houses that will lower the varability.

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?

```
sample_means50 <- rep(0, 5000)
for (i in 1:5000) {
  samp <- sample(area, 50)
  sample_means50[i] <- mean(samp)
}
hist(sample_means50, breaks = 25,main = "Area of Houses in sample_means50", x
lab = "Area")
```



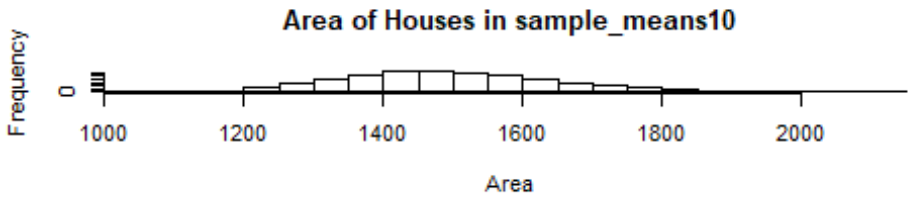

```
mean(sample_means50) #The mean sample_means50

## [1] 1498.77

sd(sample_means50) #The standard deviation of sample_means50

## [1] 70.42238
```

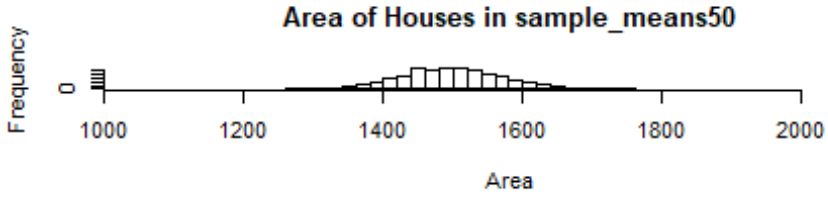
There are 5000 elemenets in sample_means50. This is a normal distribution. The mean is 1498.77 while the standard deviation is 70.42238. If we were to collect 50,000 sample means, I would expect the mean to be closer to the population mean of 1499.69 with a normal shaped distribution.

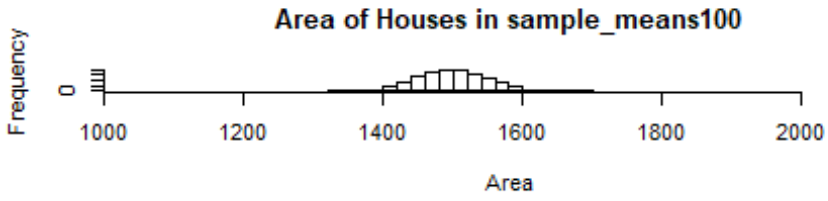
Exercise 5: When the sample size is larger, what happens to the center (mean) of the sampling distribution? What about the spread (standard deviation)?

```
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))
```

```r
xlimits = range(sample_means10)
hist(sample_means10, breaks = 20, xlim = xlimits, main = "Area of Houses in s
ample_means10", xlab = "Area")
hist(sample_means50, breaks = 20, xlim = xlimits, main = "Area of Houses in s
ample_means50", xlab = "Area")
hist(sample_means100, breaks = 20, xlim = xlimits, main = "Area of Houses in
sample_means100", xlab = "Area")
```








```r
mean(sample_means10) #The mean of sample_means10
```

```
## [1] 1497.452
```

```r
sd(sample_means10) #The standard deviation of sample_means10
```

```
## [1] 156.1729
```

```r
mean(sample_means50) #The mean of sample_means50
```

```
## [1] 1498.77
```

```r
sd(sample_means50) #The standard deviation of the area
```

```
## [1] 70.42238
```

```r
mean(sample_means100) #The mean of sample_means100
```

```
## [1] 1499.231
```

```r
sd(sample_means100) #The standard deviation of sample_means100
```

```
## [1] 49.98806
```

When the sample size is larger, the mean approaches the population mean of 1499.69. The mean of sample_means100 is 1499.231. The spread decreases as the sample size increases. The spread for sample_mean10 is 156.1729 while sample_means100 is 49.98806.

Homework Assignment

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

```
pricesamp1 <- sample(price, 50)
mean(pricesamp1) #The mean of pricesamp1
```

```
## [1] 181007.8
```

The best point estimate of the population mean for home price is 181007.8.

2. Since you have access to the population, simulate the sampling distribution for the sample mean ofhome price by taking 5000 samples from the population of size 50 and computing 5000 price samplemeans. Store these means in a vector called sample_price_means50. Plot the data, then describe the shape of this simulated sampling distribution. Based on this simulated sampling distribution, what would you guess the mean home price of the population to be?

```
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, main = "Prices of Houses in sample_price_means50",
xlab = "Price")
```


```
mean(sample_price_means50) #The mean of sample_price_means50
```

```
## [1] 180846.1
```

```
sd(sample_price_means50) #The standard deviation of sample_price_means50
```

```
## [1] 11349.31
```

The shape of the distribution can be described as a normal distribution. I would guess that the mean home price of the population would be 180800.

3. Change your sample size from 50 to 150, and then generate a simulated sampling distribution using the same method as above. Store these means in a new vector called sample_price_means150. Compare and contrast the shape, center (mean), and spread (standard deviation) of your simulated sampling distributions for n = 50 and n = 150. Based on your simulated sampling distribution for samples of size n = 150, what would you guess to be the mean sale price of homes in Ames? Finally, calculate and report the actual population mean.

```
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, main = "Price of Houses in sample_means100", xlab
= "Price")
```


```
mean(sample_price_means150) #The mean of sample_price_means150
```

```
## [1] 180791.8
```

```
sd(sample_price_means150) #The standard deviation of sample_price_means150
```

```
## [1] 6299.991
```

```
mean(price) #The mean of the whole sample
```

```
## [1] 180796.1
```

The shape of this histogram would be best described as a noraml distribution. This would be similar to the n = 50 distribution. The mean of n=50 is 180846.1 while n=150 is 180791.8. The standard deviation for n=50 is 11349.31 while n=150 is 6299.991. As the sample size increases, the standard devition decreases while the mean gets closer to the population mean. I would guess that the mean price of homes would be closer to 180800. The mean price of homes is 180796.1.

4. Of the sampling distributions from #2 and #3, which has a smaller spread (standard deviation)? If we're concerned with making estimates that are more often close to the true value, would we prefer a sampling distribution with a large or small spread? Explain your reasoning.

The variable sample_price_means150 has a smaller standard deviation of 6299.991. In order to make a good estimate, we would like distributions to have a small spread. With a small spread, we know that the data is clustered around the mean which makes the data more reliable.