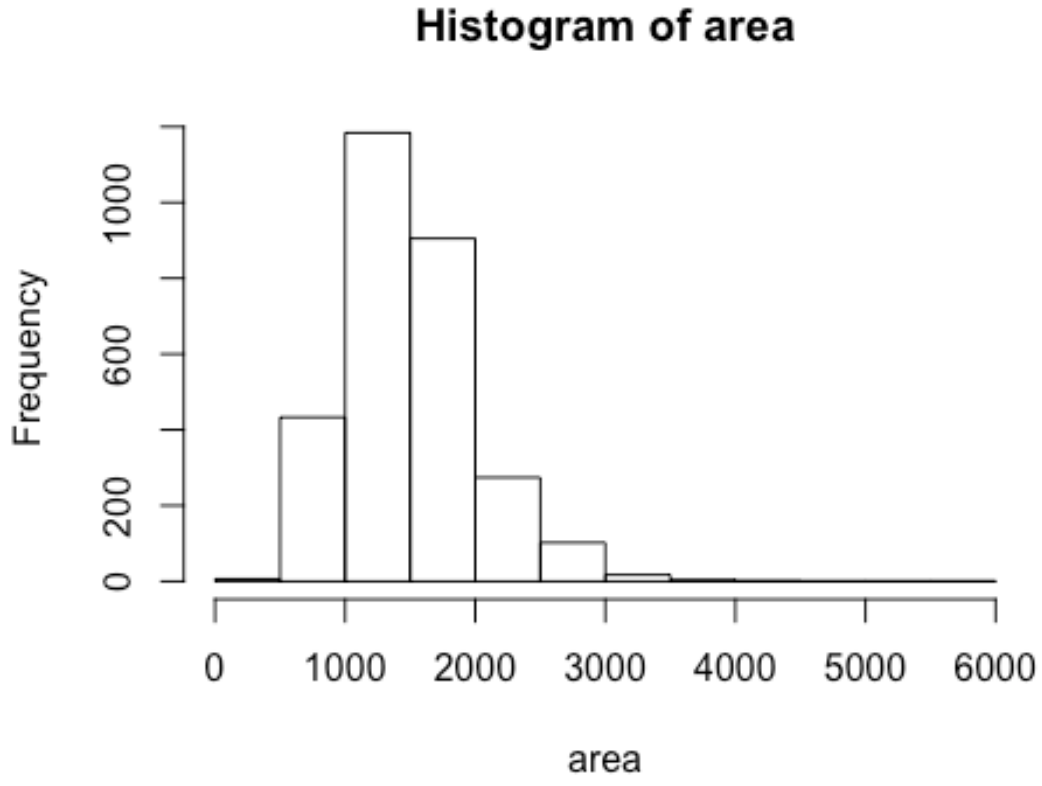
lab5andrews

Christopher Andrews

11/18/2018

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

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

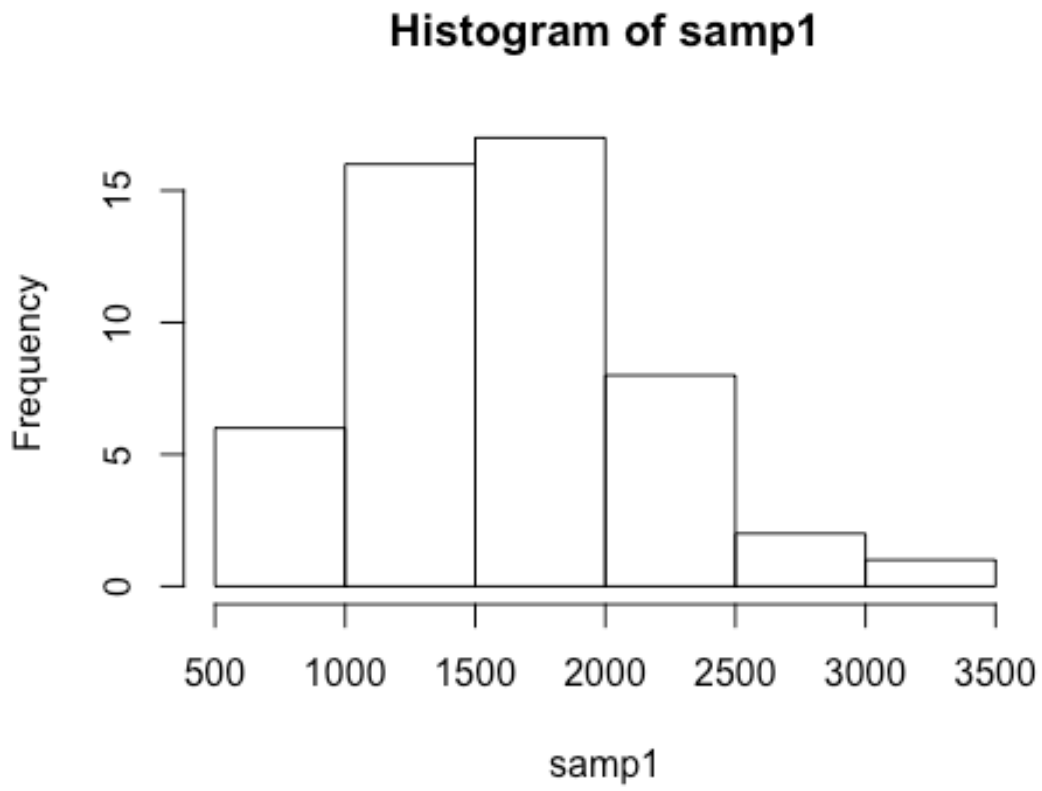
The shape of this histogram is right skewed and unimodal. The center is at 1500 and its spread (standard deviation) is around 505.5

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

➡️ **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 spread of this set of data is not as right-skewed as above and is more uniformly distribued across the x-axis. The mean is only 10 higher then the above data set, lying at 1614. The standard deviation of this set is about 50 lower than the above data sample, and its value is 558.3744. Its shape other than being more uniformly distributed across the x-axis is also bimodal.

```
summary(samp1)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     810    1142    1614    1635    1970    3390

sd(samp1)

## [1] 558.3744

hist(samp1)
```




```
mean(samp1)
```

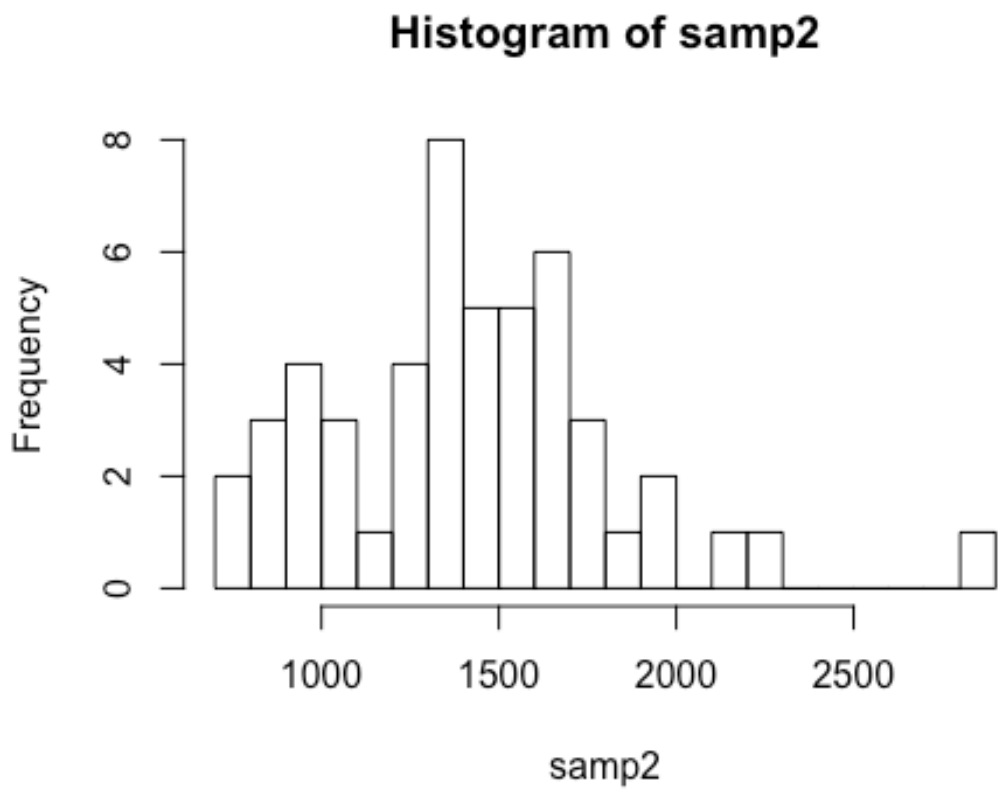
```
## [1] 1635.02
```

➡️ **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?**

This sample is like the first one in that it is right skewed. The median is much lower and the mean is about the same, and the standard deviation is much, much lower. As the size of the sample gets higher it is my belief that the standard deviation would get much lower and the spread shape would become much more defined and vary less. I think the more accurate estimate of the population mean would be the one with 1000 as its sample size. After doing the calculations, you can see that the standard deviation gets larger between the sample size of 50 to the sample size of 100, then has a slight increase when it goes from 100 to 1000. And as seen when the sample size is 5000, it gets much smaller.

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

sampW100 <- sample(area, 100)
sampW1000 <- sample(area, 1000)
```

```r
hist(samp2, breaks = 25)
```




```r
summary(samp2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     784    1186    1410    1420    1626    2822
```

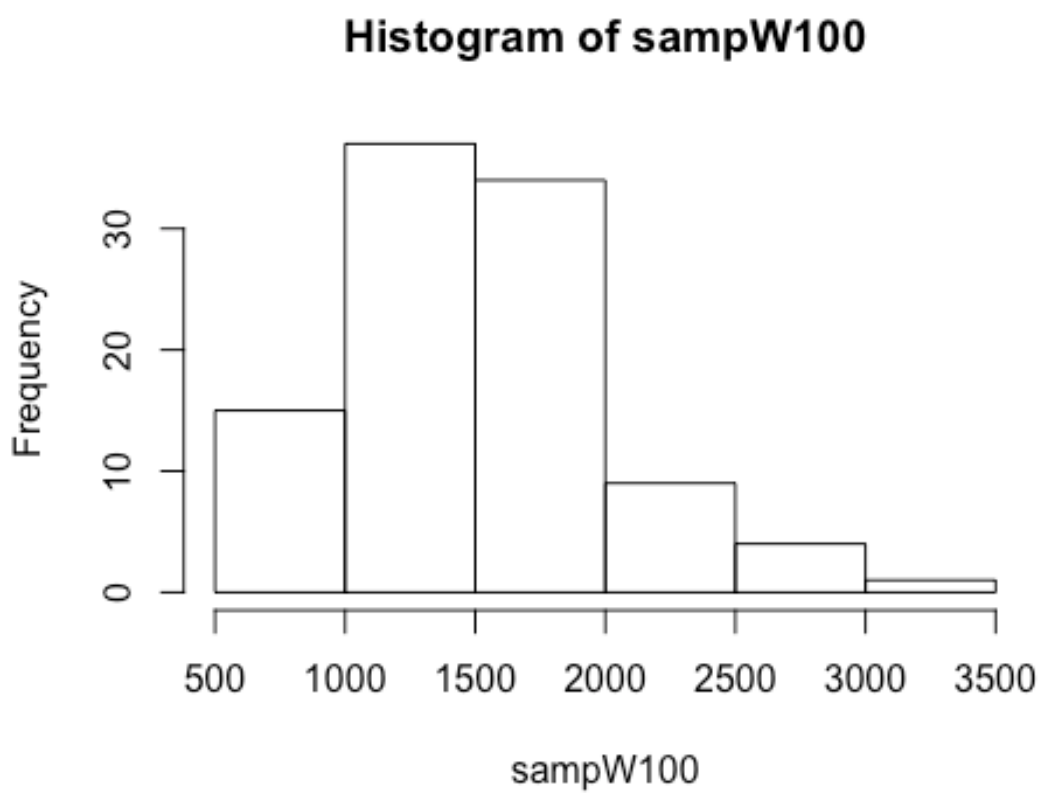
```r
print("50")
```

```
## [1] "50"
```

```r
sd(samp2)
```

```
## [1] 406.4352
```

```r
hist(sampW100)
```




```
summary(sampW100)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     520    1117    1439    1499    1730    3500
```

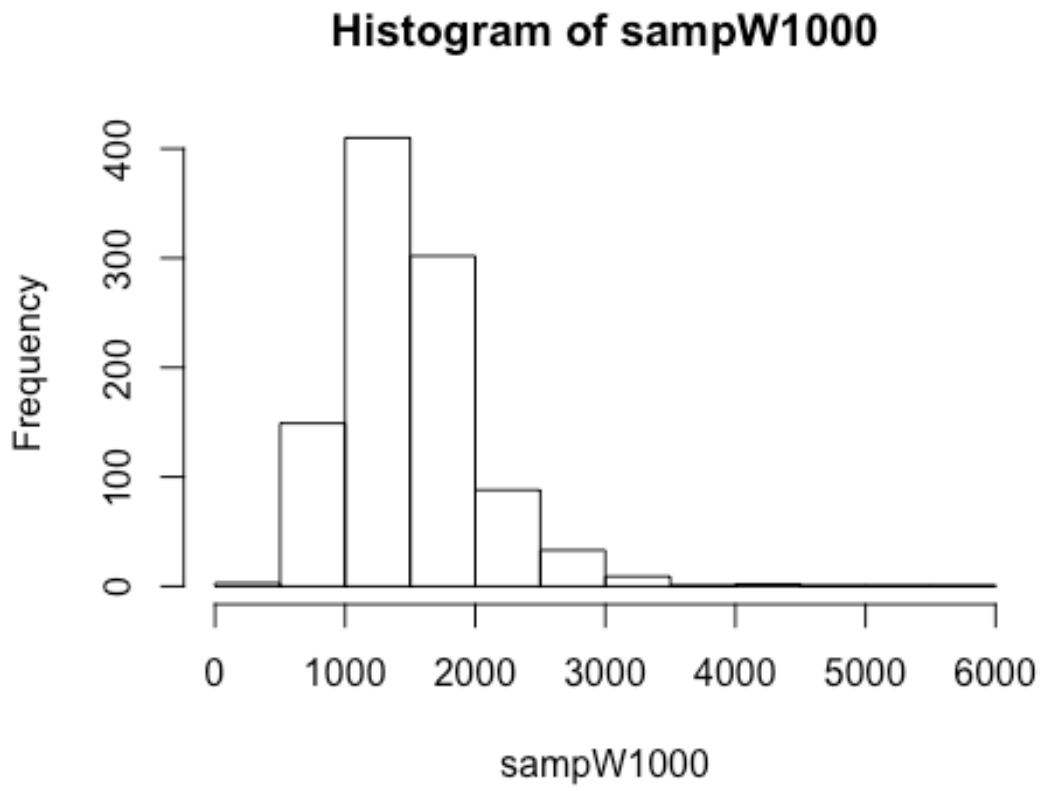
```
print("100")
```

```
## [1] "100"
```

```
sd(sampW100)
```

```
## [1] 517.806
```

```
hist(sampW1000)
```




```
summary(sampW1000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     407    1124    1432    1508    1742    5642
```

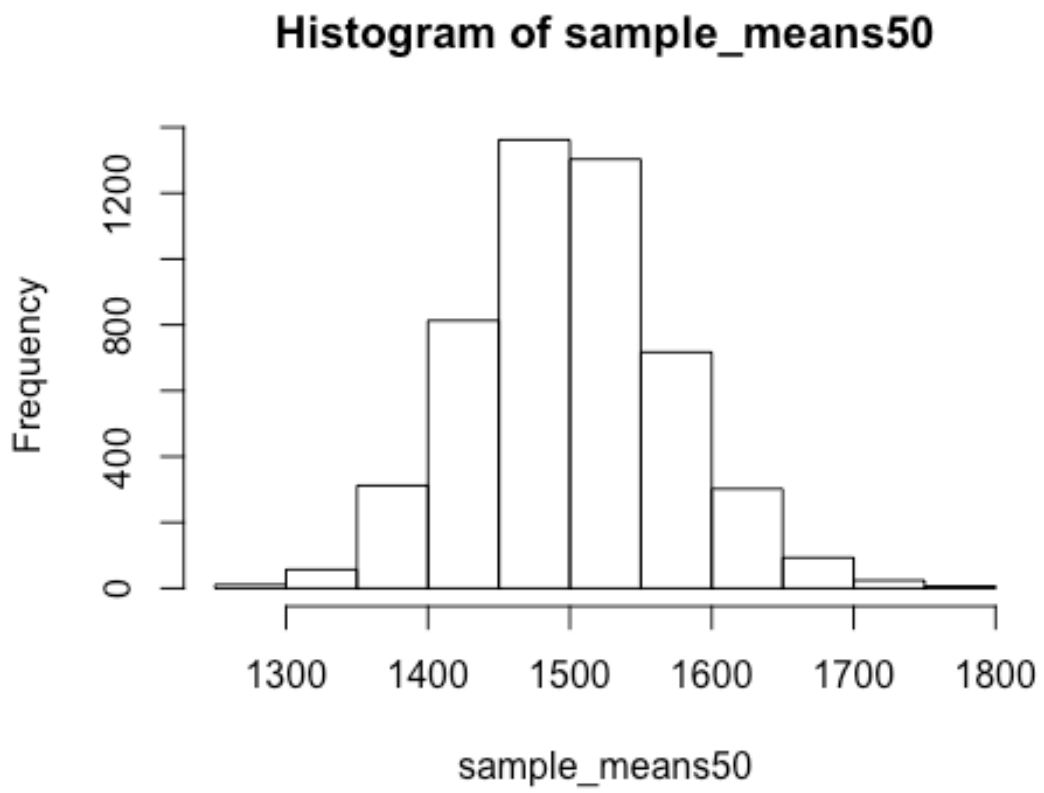
```
print("1000")
```

```
## [1] "1000"
```

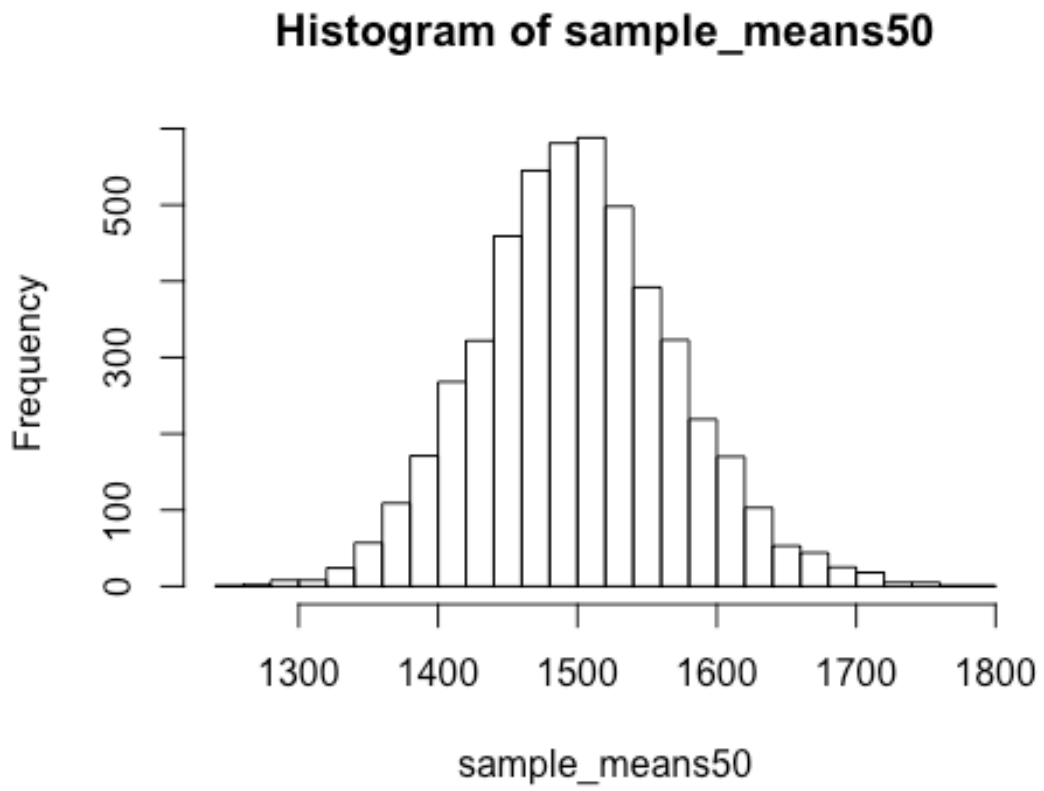
```
sd(sampW1000)
```

```
## [1] 548.7321
```

```
sample_means50 <- rep(0, 5000)
for (i in 1:5000) {
samp <- sample(area, 50)
sample_means50[i] <- mean(samp)
}
hist(sample_means50)
```




```r
hist(sample_means50, breaks = 25)
```

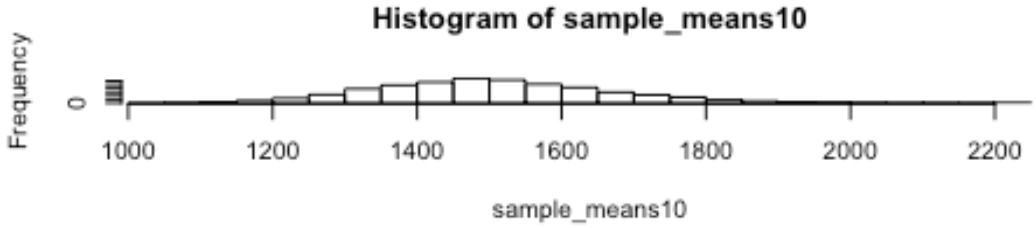


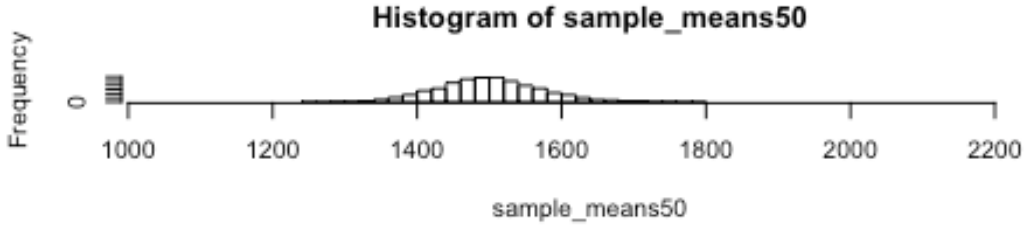
➡️ **Exercise 4: How many elements are there in sample_means50? Describe the shape, center (mean), and spread (standard deviation) of the sampling distribution. How would you expect the sampling distribution to change if we instead collected 50,000 sample means?**

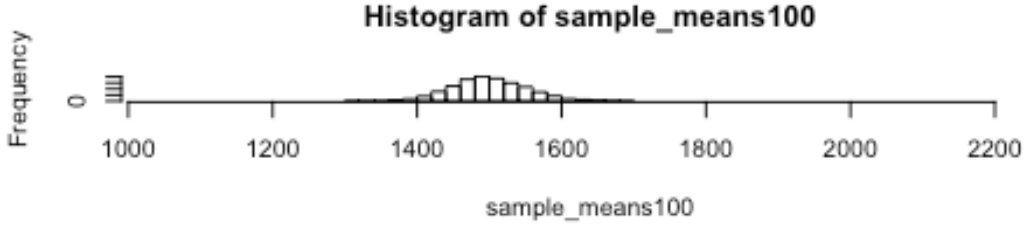
In sample_mean50 there is 5000 sampled elements. The spread of this distribution is uniform and symmetrical. The mean and medians are very similar to the earlier samples, with the center being at 1500. As you could have guessed with a larger sample size that the standard deviation is much, much lower. With a collection of sample means 10 times higher then 5000, the median and means are likely to stay similar to what they are now, with some deviation , the spread will become more defined around the center of the graph and will become even more symmetrical, and the standard deviation should and most likely will shrink even more in size.

```
summary(sample_means50)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1251    1453    1498    1500    1545    1790

sd(sample_means50)

## [1] 71.46894
```

```r
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)
```







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

The mean moves slightly positively and negatively as the sample size changes. As the amount of samples increases than the mean gets closer and closer to its true mean. And the shapes, while all being unquestionably uniform and symmetrical, grow inwards towards the mean as the sample size gets higher. It grows steeper and steeper, and the spread grows smaller and smaller. The standard deviation grows smaller and smaller as the sample size increases.

```
summary(sample_means10)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1038    1393    1495    1504    1605    2240
```

```
sd(sample_means10)
```

```
## [1] 160.4187
```

```
summary(sample_means50)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1251    1453    1498    1500    1545    1790
```

```
sd(sample_means50)
```

```
## [1] 71.46894
```

```
summary(sample_means100)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1319    1466    1498    1500    1533    1682
```

```
sd(sample_means100)
```

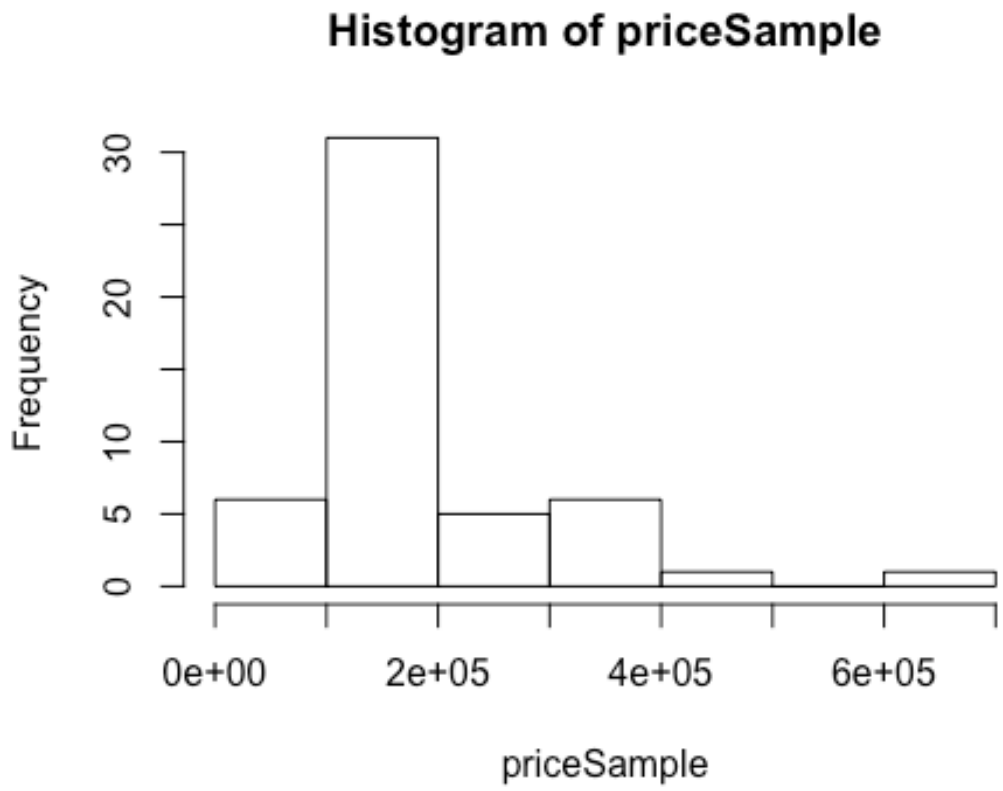
```
## [1] 50.53591
```

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?

The distribution is right skewed and unimodal. To the right of the mean, it falls off extremely. The best point estimate of the mean with a sample of 50 is 189893.

```
priceSample <- sample(price, 50)
hist(priceSample)
```

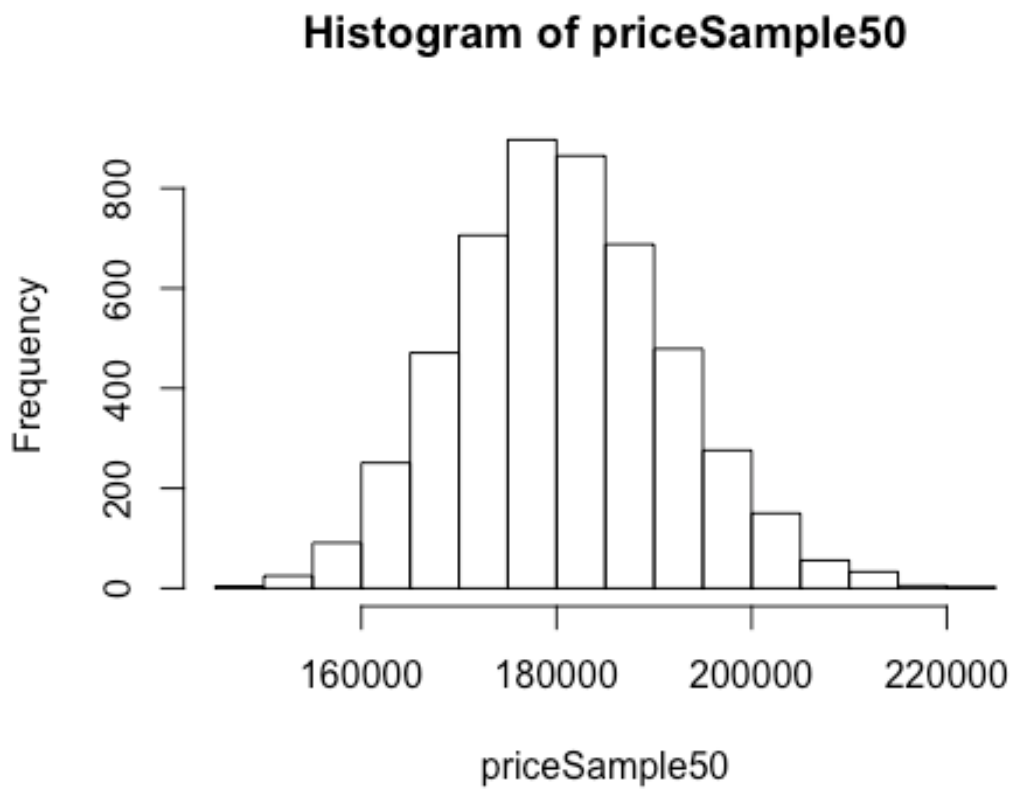



```
summary(priceSample)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68104  126250  155200  189893  209000  610000

sd(priceSample)

## [1] 104698.8
```

2. **Since you have access to the population, simulate the sampling distribution for the sample mean of home price by taking 5000 samples from the population of size 50 and computing 5000 price sample means. 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?**

   The shape of this distribution is undoubtedly uniform, unimodal and symmetrical. The sampling mean is 180745, which is lower than the previous sampled mean.

```r
priceSample50<-rep(0,5000)
for(i in 1:5000){
samp<-sample(price,50)
priceSample50[i]<-mean(samp)
}

hist(priceSample50, breaks = 25)
```




```r
summary(priceSample50)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  146279  173153  180389  180745  187991  223928
```

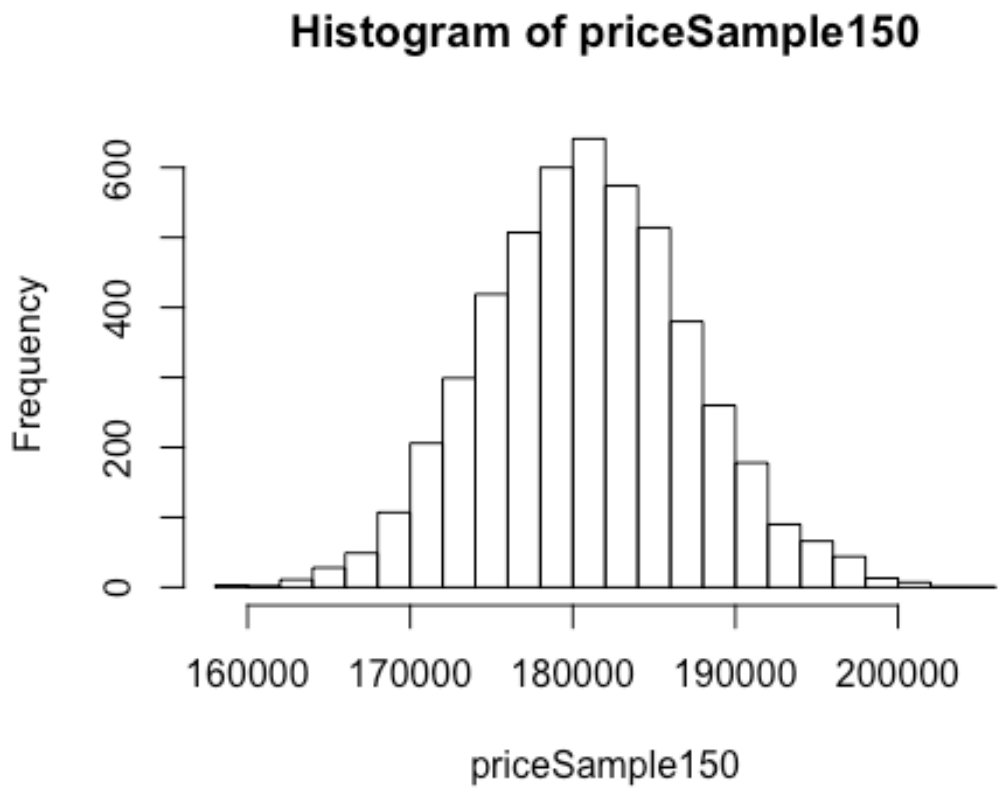
```r
sd(priceSample50)
```

```
## [1] 11141.67
```

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.**

The spreads a distributions are similar and the shape is uniform, unimodal and symmetrical like the other sample if 50. The only difference is that the 150 shape is steeper tan the one of 50. The standard deviation is just about cut in half, as you could guess it is much smaller as there is a higher sample. Based on the sample of 150 the population mean is 180903, and the actual population mean is 180796.1.

```
priceSample150<-rep(0,5000)
for(i in 1:5000){
  samp<-sample(price,150)
  priceSample150[i]<-mean(samp)
}

hist(priceSample150, breaks = 25)
```



```
summary(priceSample150)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  159725  176548  180802  180903  185154  204102

sd(priceSample150)

## [1] 6411.417
```

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.**

Out of the two, the sampling distributino with the higher number if n (n = 150) has the smaller spread and standard deviation. If we were converned with making estimates that are more often close to the true value, we would surely use the sampling distributino where n = 150.