# Lab 5: Sampling Distributions

*Quentin Terry*

*November 18, 2018*

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

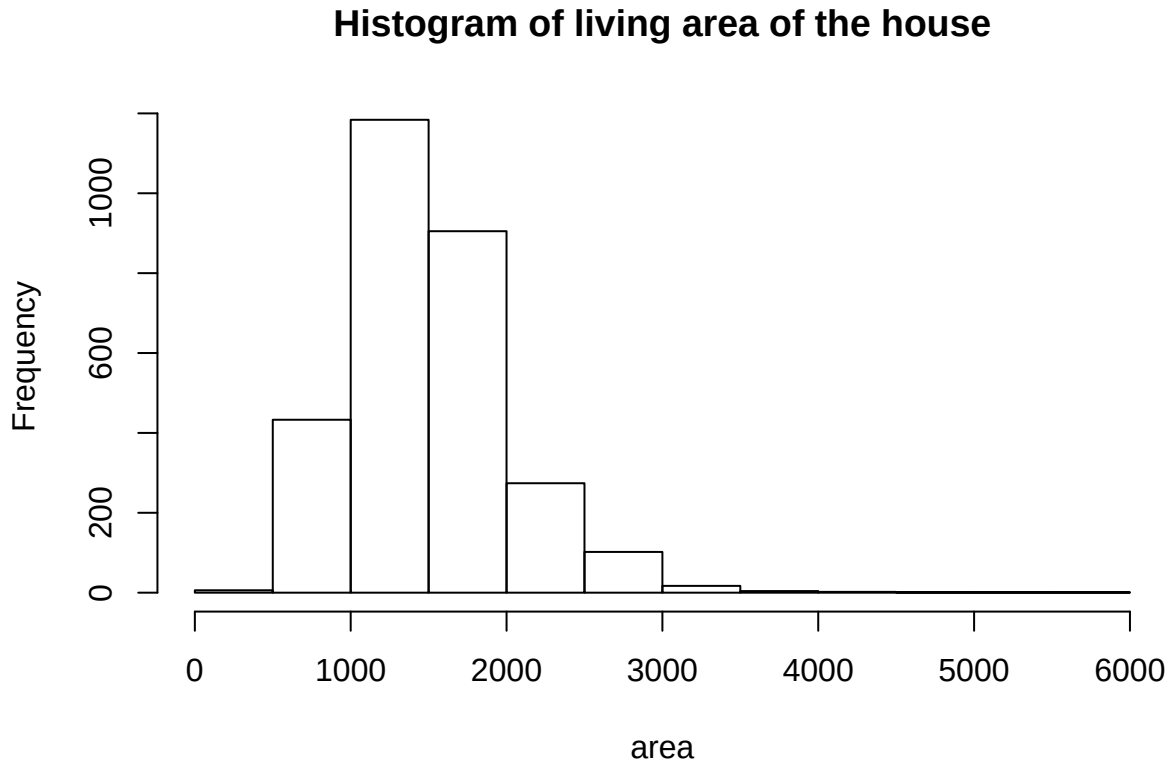
```
area <- ames$Gr.Liv.Area
price <- ames$SalePrice
```

**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 = "Histogram of living area of the house")
```




```
sd(area)
```

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

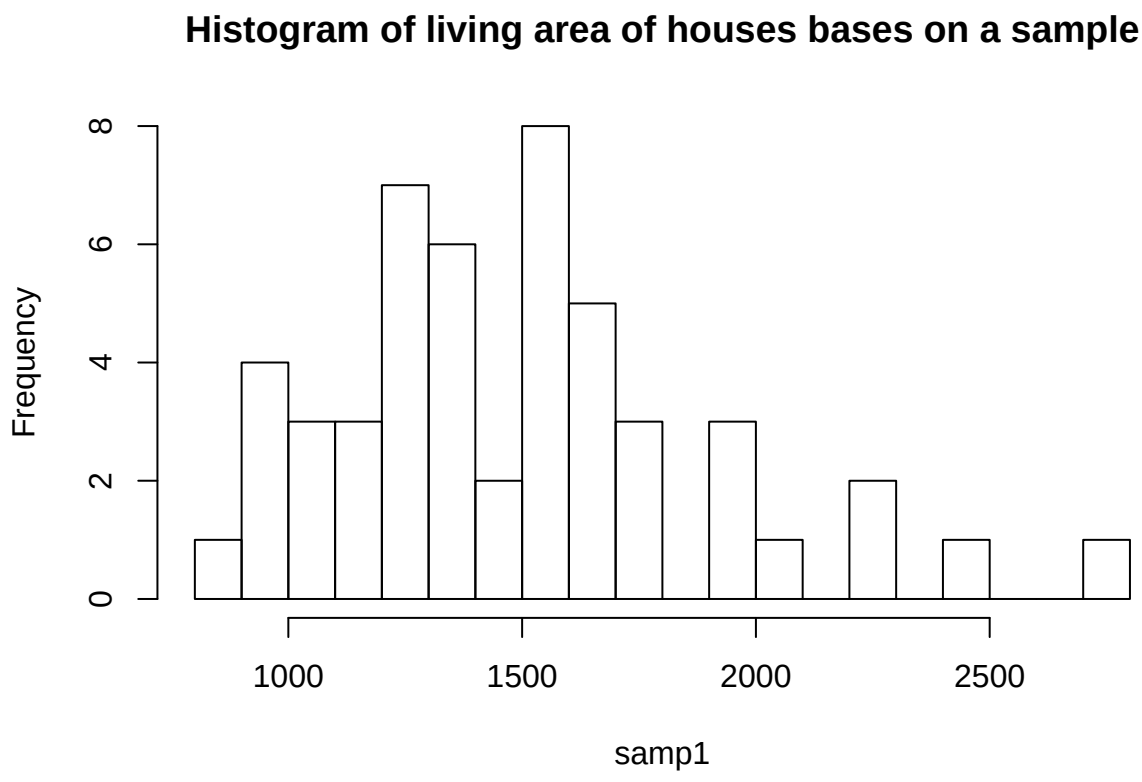
-The Shape of this distributon is right skewed. The center of the distribution is 1500 and the standard deviation is 505.5

**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?**

```
set.seed(001)

samp1 <- sample(area, 50)

summary(samp1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     864    1218    1463    1491    1679    2730
```

```
hist(samp1, main = "Histogram of living area of houses bases on a sample", breaks = 20)
```




```
sd(samp1)
```

```
## [1] 401.2159
```

-The histogram of the sample is also right-skewed, The center is 1491 and the standard deviation is 401.2. In comparison the sample mean is lower than the mean in the population distribution in exercise 1.the standard deviation is also lower, however, they are both right-skewed.

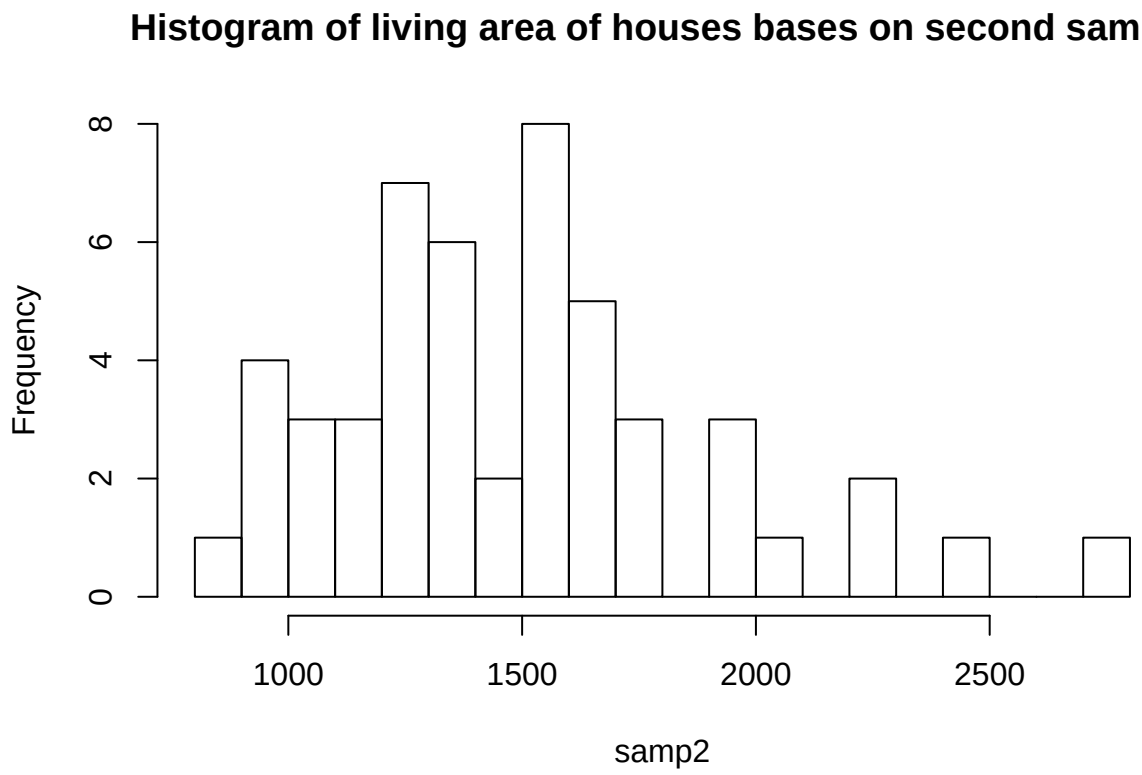
**Exercise 3: Take a second sample, also of size 50, and name it samp2. How does the mean of samp2 compare with the mean of samp1? Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population mean? Why?**

```
set.seed(001)
samp2 <- sample(area,50)

samp100 <- sample(area, 100)

samp1000 <- sample (area, 1000)
```

```
set.seed(001)
summary(samp2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     864    1218    1463    1491    1679    2730
```

```
hist(samp2, main = "Histogram of living area of houses bases on second sample", breaks = 20)
```




```
sd(samp2)
```

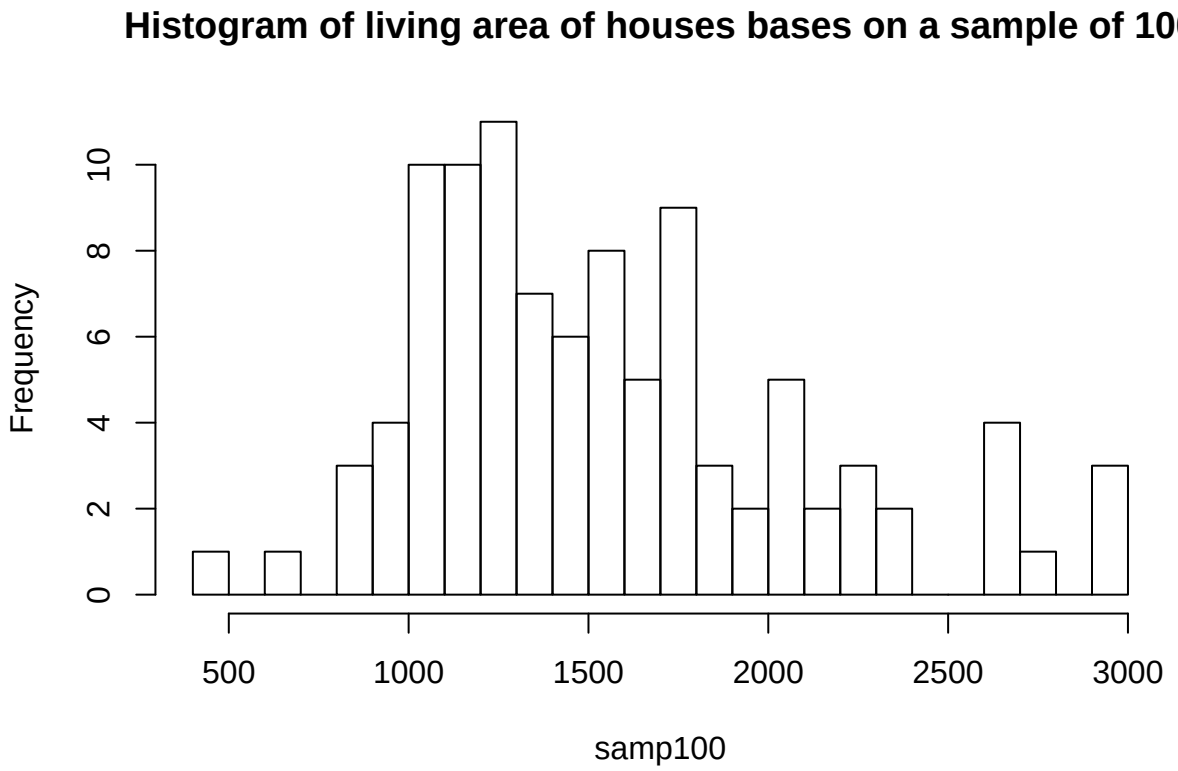
```
## [1] 401.2159
```

```
summary(samp100)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##     492    1178    1442    1557    1806    2956
```

```
hist(samp100, main = "Histogram of living area of houses bases on a sample of 100", breaks = 20)
```



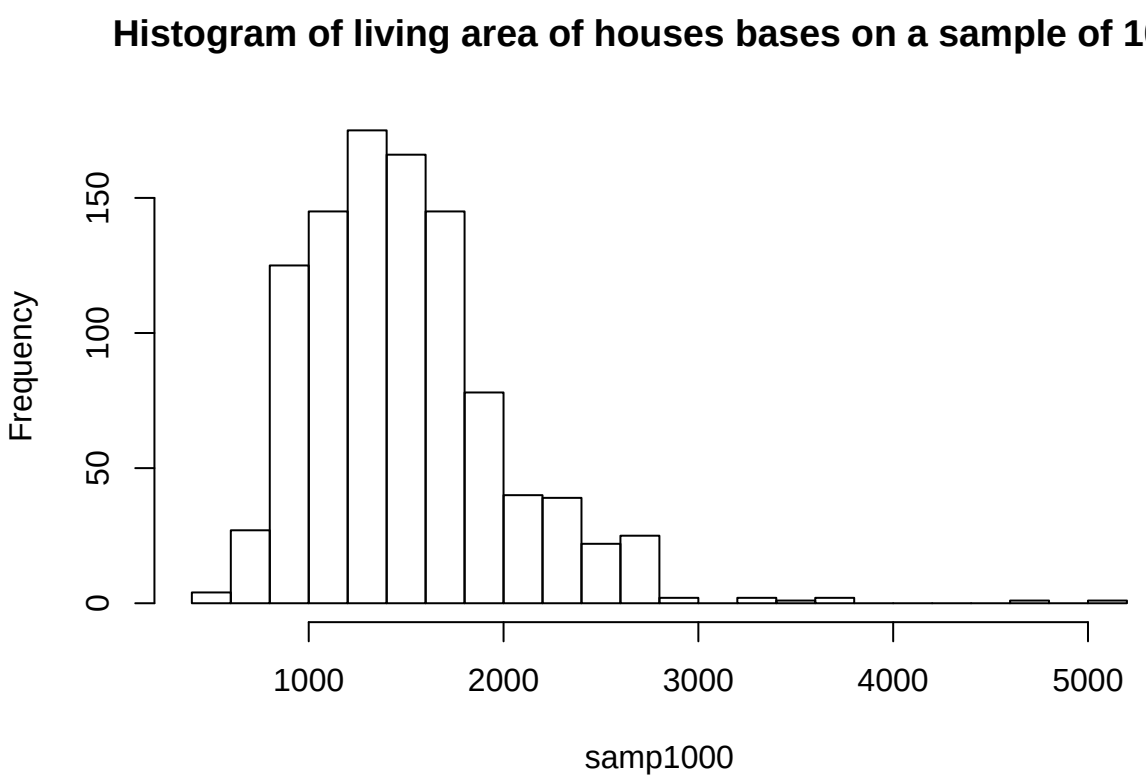

```
sd(samp100)
```

```
## [1] 533.3047
```

```
summary(samp1000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     407    1128    1430    1492    1728    5095
```

```
hist(samp1000, main = "Histogram of living area of houses bases on a sample of 1000", breaks = 20)
```



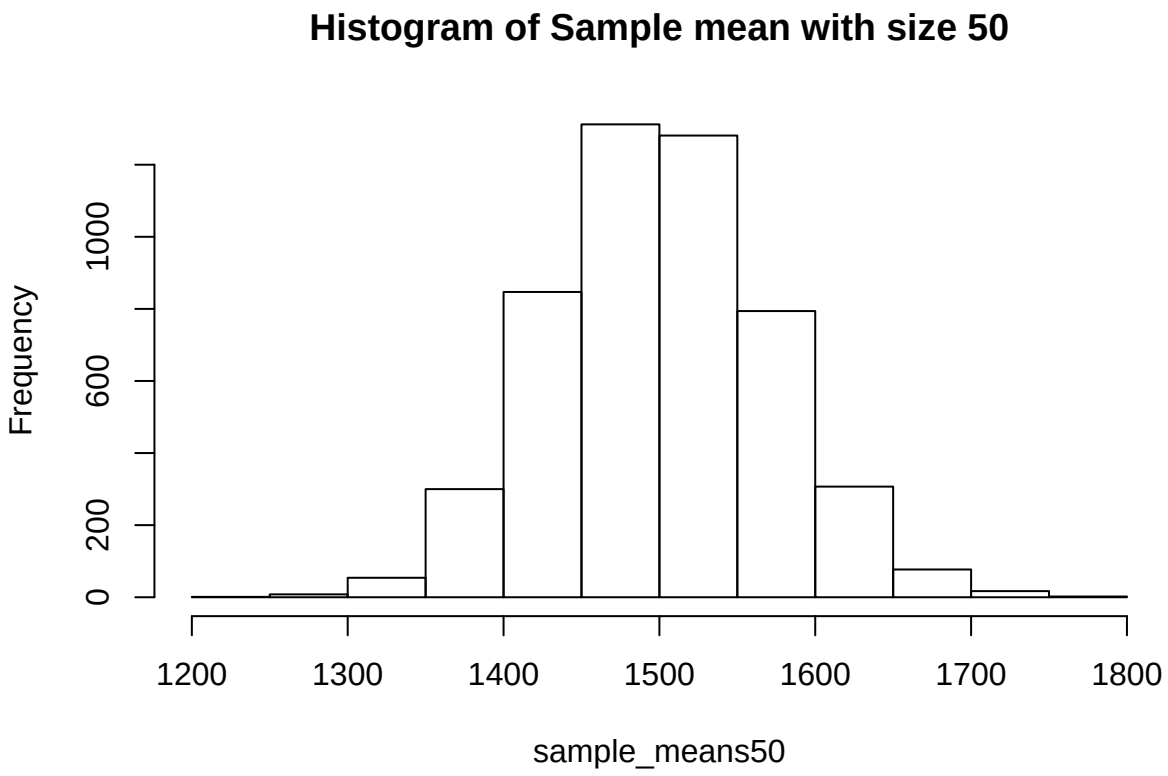

```
sd(samp1000)
```

```
## [1] 505.5156
```

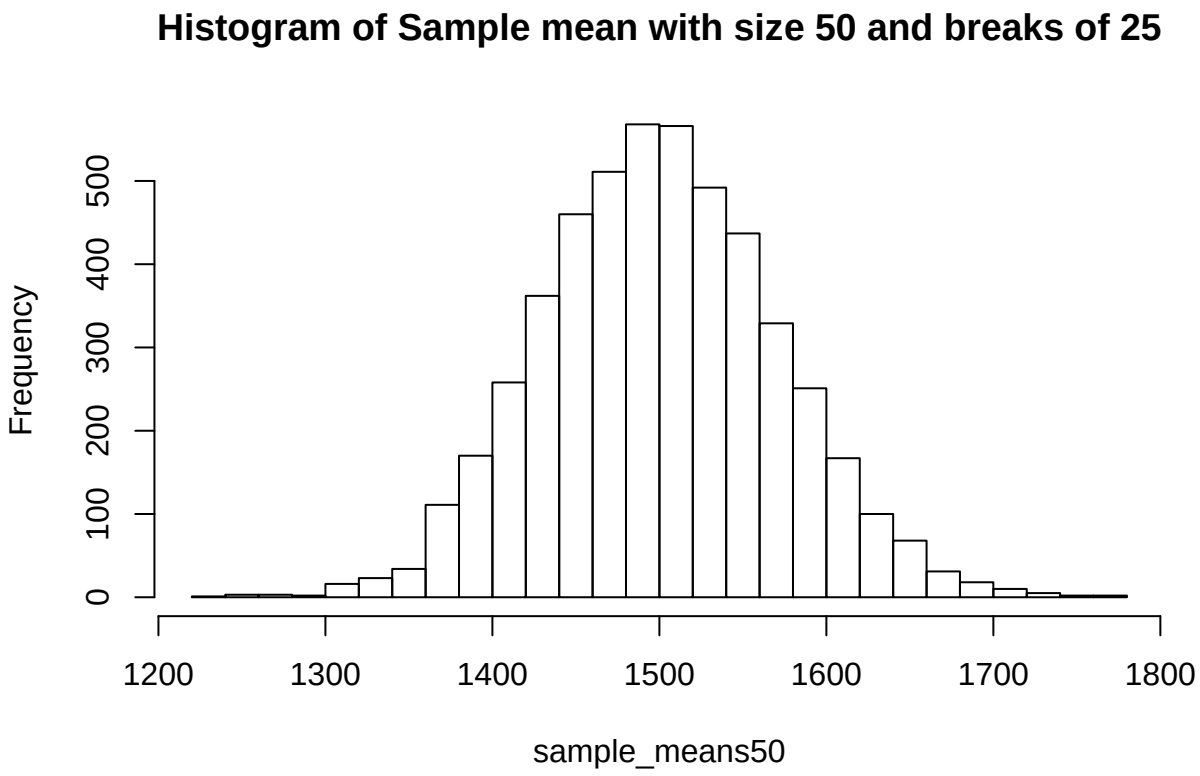
The second sample of size 50 is similar to the first sample, however the means are slightly different due to them being different samples and not the same seed. I think that the size of 1000 would provide a more accurate estimate for the mean because there are more samples to be summarized.

**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?**

```
set.seed(001)
sample_means50 <- rep(0, 5000)

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

hist(sample_means50, main = "Histogram of Sample mean with size 50")
```




```r
hist(sample_means50, breaks = 25, main = "Histogram of Sample mean with size 50 and breaks of 25")
```




```
summary(sample_means50)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1235    1452    1499    1501    1547    1766
```

```
sd(sample_means50)
```

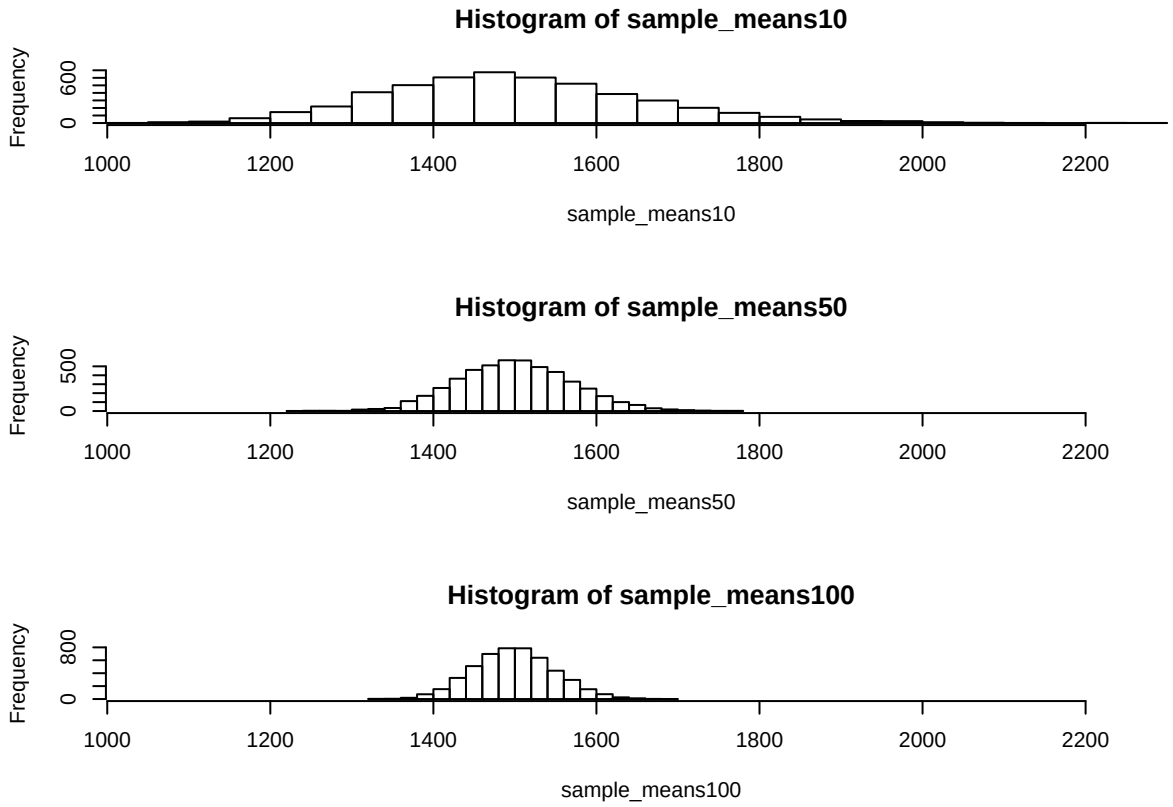
```
## [1] 70.51946
```

There are 5000 elements in the sample_means50. The shape for this sampling distribution is uniform and symmetric. The center is 1501 which is similar to the previous sample means. The standard deviation however is much lower than the previous sample distributions at 70.519. If 50,000 sample means were collected instead, the standard deviation would be even smaller.

### 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)
```

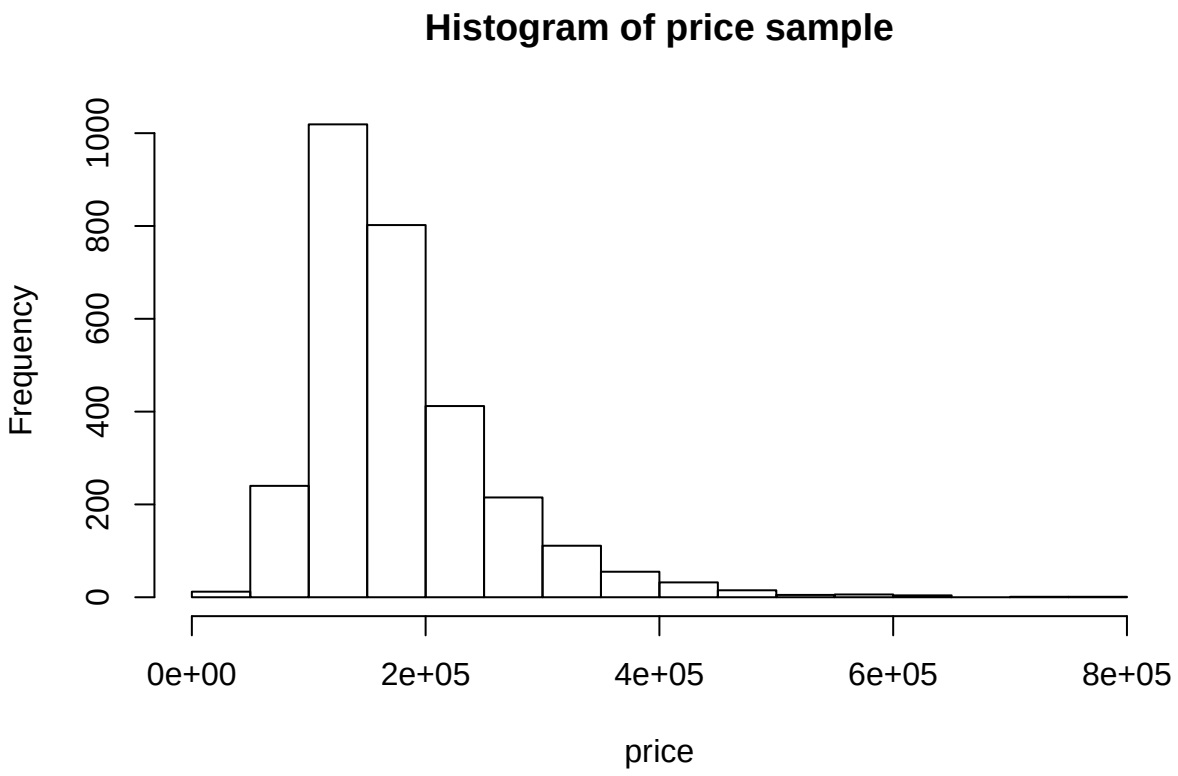
```
}

par(mfrow = c(3, 1))

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)
```




-When the sample size is larger, the mean stays the same dispite how large the sample size is. However, the spread of the sampling distribution gets smaller as the sample size increases.

## 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?

```
pricesamp <- sample(price, 50)

hist(price, breaks = 20, main = "Histogram of price sample")
```




```
summary(price)
```

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

```
sd(price)
```

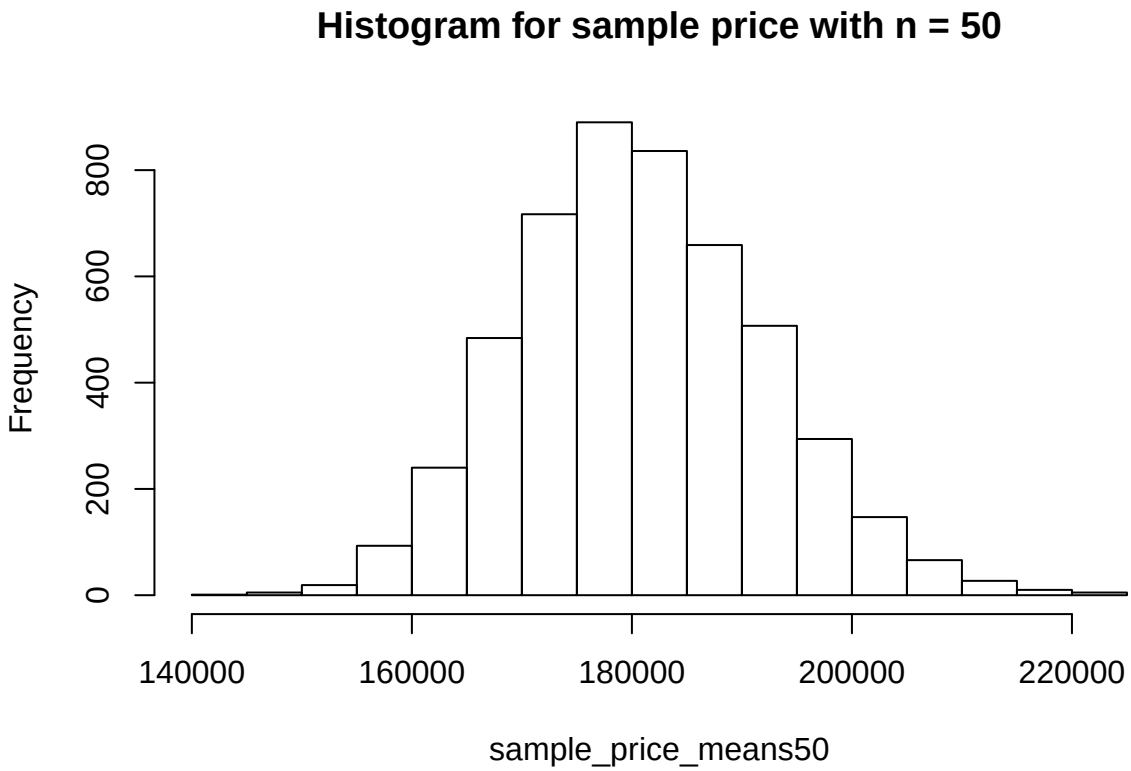
```
## [1] 79886.69
```

- This population distribution is right-skewed and the ebst point estimate of the population is 180796

**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?**

```
set.seed(001)
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, breaks = 20, main = "Histogram for sample price with n = 50")
```




```
summary(sample_price_means50)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  142663  173051  180294  180867  188314  224643
```

```
sd(sample_price_means50)
```

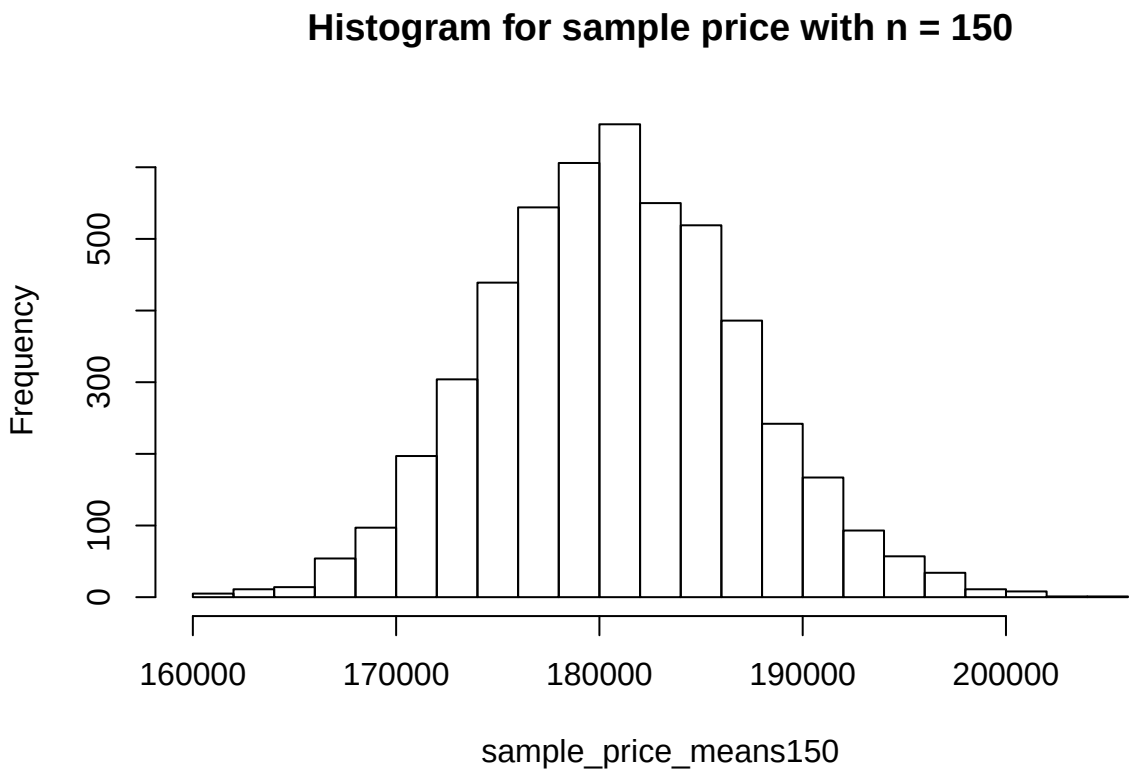
```
## [1] 11239.13
```

-The shape of this population distribution is uniform and symmetrical. The mean home price of the population would then be 180867.

**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.**

```
set.seed(001)
sample_price_means150 <- rep(0,5000)
```

```r
for(i in 1:5000){
samp<-sample(price,150)
sample_price_means150[i]<-mean(samp)
}

hist(sample_price_means150, breaks = 20, main = "Histogram for sample price with n = 150")
```




```r
summary(sample_price_means150)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  160798  176505  180728  180823  184994  205015
```

```r
summary(price)
```

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

```r
sd(sample_price_means150)
```

```
## [1] 6272.882
```

- Again the shape of this distribution is uniform and symmetrical. The center is 180823, which is very similar to the previous sample. The spread is 6273 which is much lower than the sampling distribution for n = 50. The mean sale price for homes in Ames would be 180823. The actual population mean is very similar at 180796.

**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 sampling distribution with n = 150 has a smaller spread. If we are concerned with making estimates we wouild perfer to have the sampling distribution of n = 150 due to the fact that a smaller spread means that it is closer to the true value.