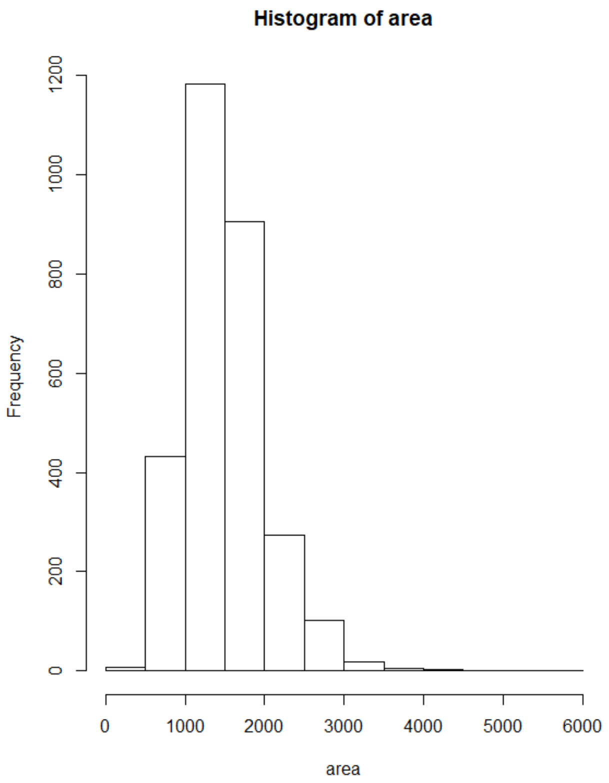
Name: Samuel L. Peoples

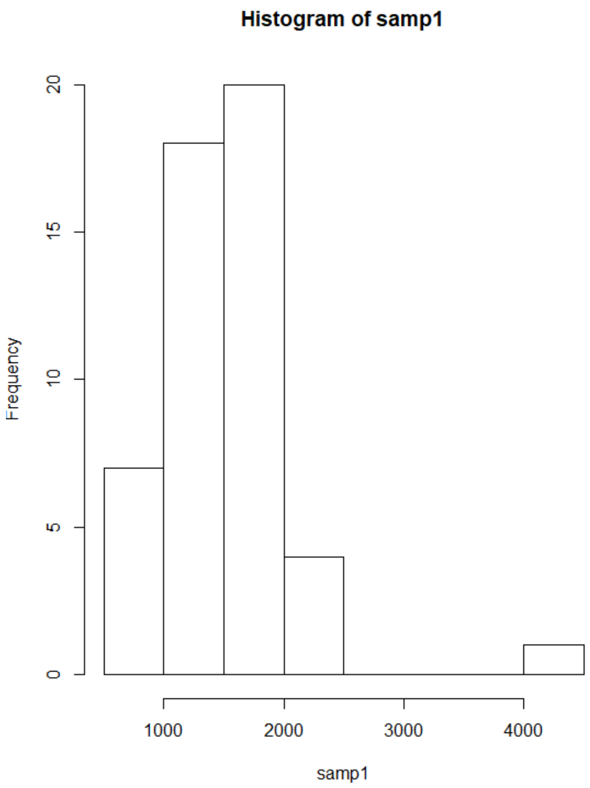
R Lab 4: Sampling Distributions

Exercise 1: Describe this population distribution.

```
> 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)
```

This distribution is skewed right, with an IQR of 617. We would expect the ground living area to be 1500 square feet.





Exercise 2: Describe the distribution of this sample. How does it compare to the distribution of the population?

```
> summary(samp1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    788    1198    1506    1544    1778    4476 
> hist(samp1)
```

In this sample of fifty files, we can see that the distribution is less obviously skewed, where the maximum of 4476 could be construed as an outlier. The IQR has been reduced to 580, but the sample behaves generally the same as the population of Ames.

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?

```
> mean(area)
                                                                     1499.69
> mean(samp1)
                                                                      1544.2
> samp2 <- sample(area, 50)
> mean(samp2)
                                                                      1560.9
> samp3 <- sample(area, 100)
> mean(samp3)
                                                                     1478.06
> samp4 <- sample(area, 1000)
> mean(samp4)
                                                                     1503.56
```

The mean area is 1499.69, where samp1 was 44.51 away from the true mean, samp2 was 61.21, samp3 was 21.63, and samp4 was only 3.87. As the sample grew in size, it became more representative of the entire population, and there was a more diverse set of data to choose from. This would be more representative of the true mean.

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)
> summary(sample_means50)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1241    1452    1497    1501    1548    1757
```

The for loop creates 5000 elements. The data is normally distributed, with a mean of 1501, and median of 1497; the standard deviation is 70.319. I believe that the distribution would be changed because there would be more samples that are close to the mean, decreasing the standard deviation.

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)}
```

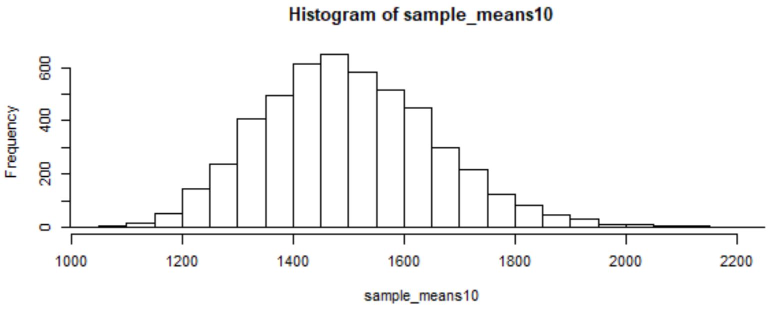
There are 100 elements which individually represent the mean of a sample of size 50 files.

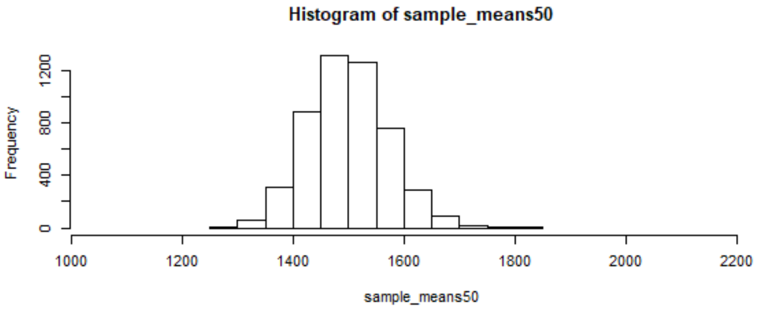
Exercise 6: When the sample size is larger, what happens to the center? What about the spread?

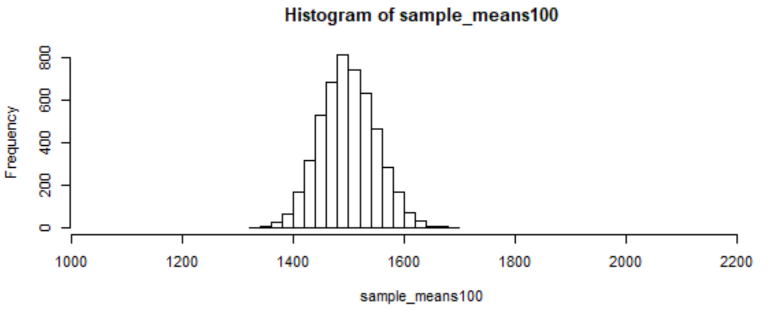
```
> hist(sample_means10, breaks = 20, xlim = xlimits)
> hist(sample_means50, breaks = 20, xlim = xlimits)
> hist(sample_means100, breaks = 20, xlim = xlimits)
```

```
> mean(sample_means10)

                                      1501.115
> sd(sample_means10)
                                      158.8582
> mean(sample_means50)
                                      1499.304
> sd(sample_means50)
                                      70.78465
> mean(sample_means100)
                                       1498.77
> sd(sample_means100)
                                      49.79445
```







Observe that the mean stays roughly near the true mean of 1499.69, with the largest difference of 1.42. Interestingly, as the samples grow in size, the spread is decreased from a standard deviation of 158.86, to a standard deviation of 49.79. This verifies our hypothesis from number 3.

On Your Own:

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

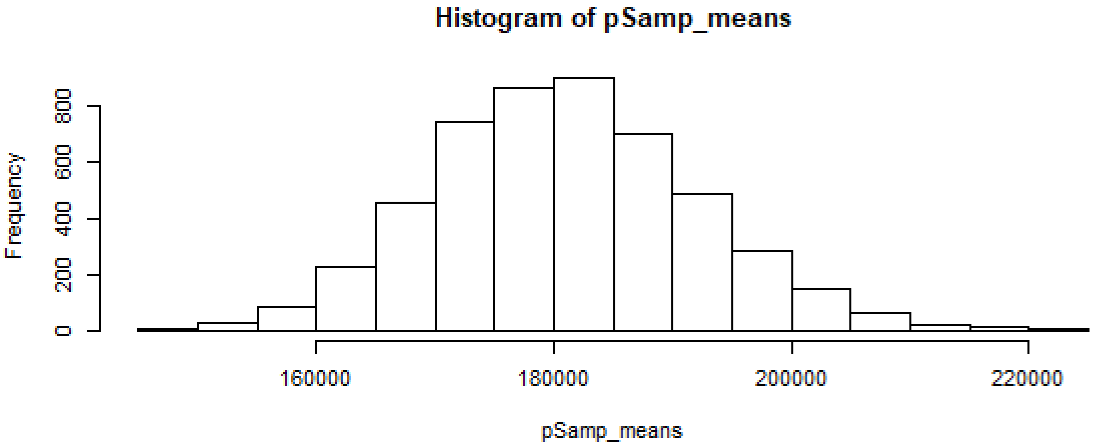
```
> pSamp1 <- sample(price, 50)
> mean(pSamp1)
175981.7
```

We can estimate that the mean is roughly 175981.7.

2. Since you have access to the population, simulate the sampling distribution for x¯pricex¯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.

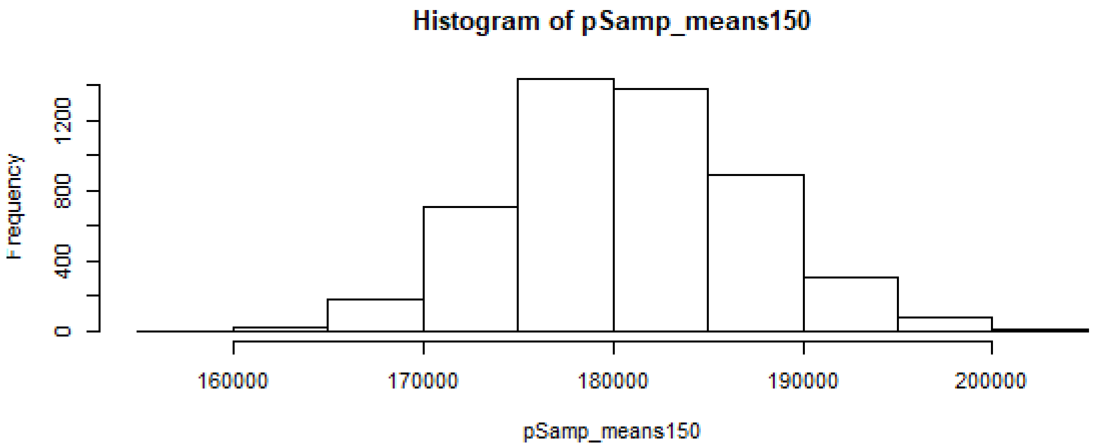
```
> pSamp_means <- rep(NA, 5000)
> for (i in 1:5000){
+     samp <- sample(price, 50)
+     pSamp_means[i] <- mean(samp)}
> hist(pSamp_means)
> summary(pSamp_means)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 147703  173242  180571  180911  188117  221900 
> sd(pSamp_means)
                                                                         11069.75
> mean(price)
                                                                         180796.1
```



This data is normally distributed with a center at 180571, mean of 180911, IQR of 14875 and Standard deviation of 11069.75. We can estimate that the mean is roughly 180911 from this distribution, where the true mean of the population price is 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?

```
> pSamp_means150 <- rep(NA, 5000)
> for (i in 1:5000){
+     samp <- sample(price, 150)
+     pSamp_means150[i] <- mean(samp)}
> 
> hist(pSamp_means150)
> summary(pSamp_means150)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 159612  176445  180522  180818  185101  202586 
> sd(pSamp_means150)
                                                                         6380.506
```



This distribution has more samples at the suspected mean and still appears to be normal; the distribution's spread is reduced. Based on this data we can estimate the mean at 180818.

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 larger sample has a smaller spread, and if we were concerned with making estimates that were more often close to the true value, we would be most interested in distributions with smaller spreads, that way the data is more representative of the expected values.