Lab4a

Bin Lin

2016-10-8

```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")
library(IS606)
```

```
##
## Welcome to CUNY IS606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='IS606') will list the demos that are available.
```

```
##
## Attaching package: 'IS606'
```

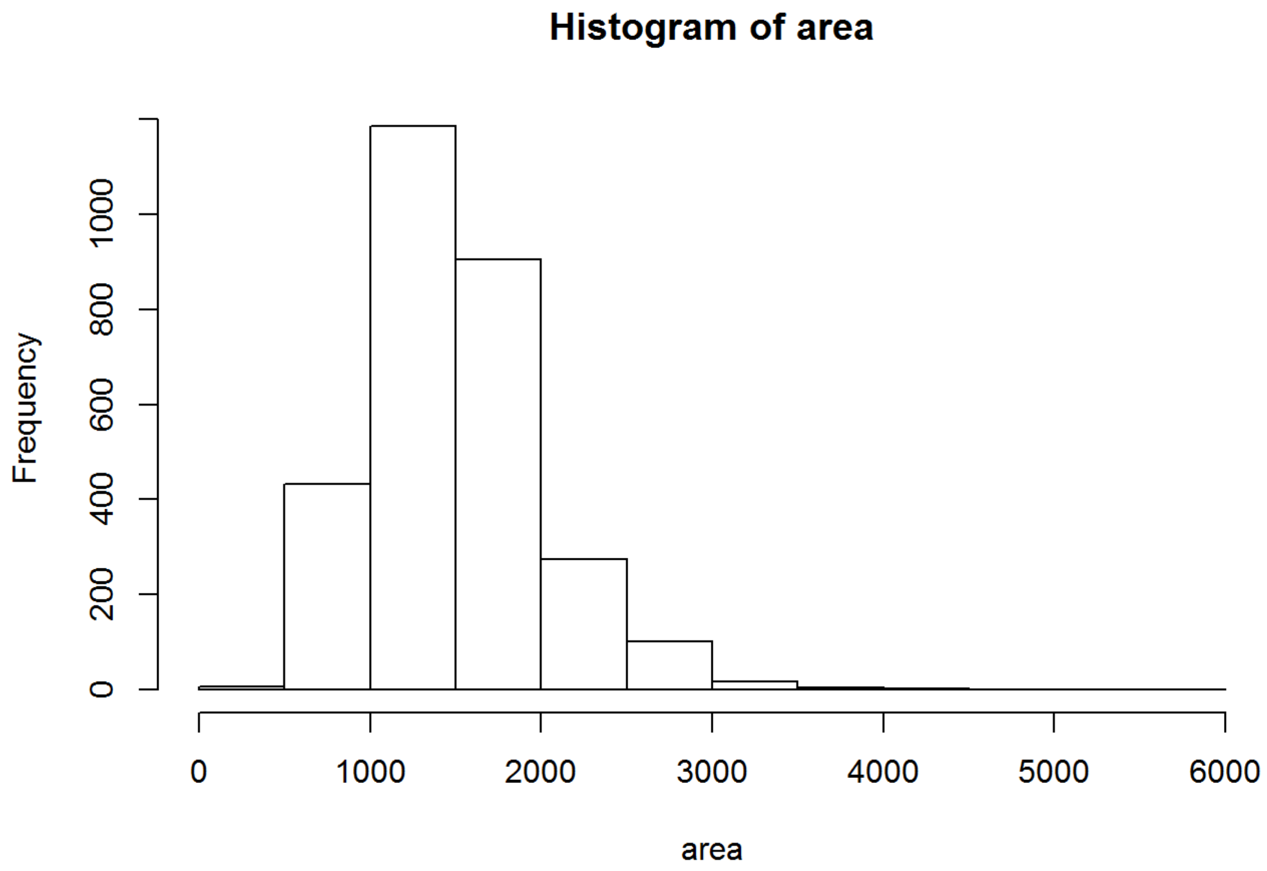
```
## The following object is masked from 'package:utils':
##
##     demo
```

```
load("ames.RData")
set.seed(10)
```

```
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)
```

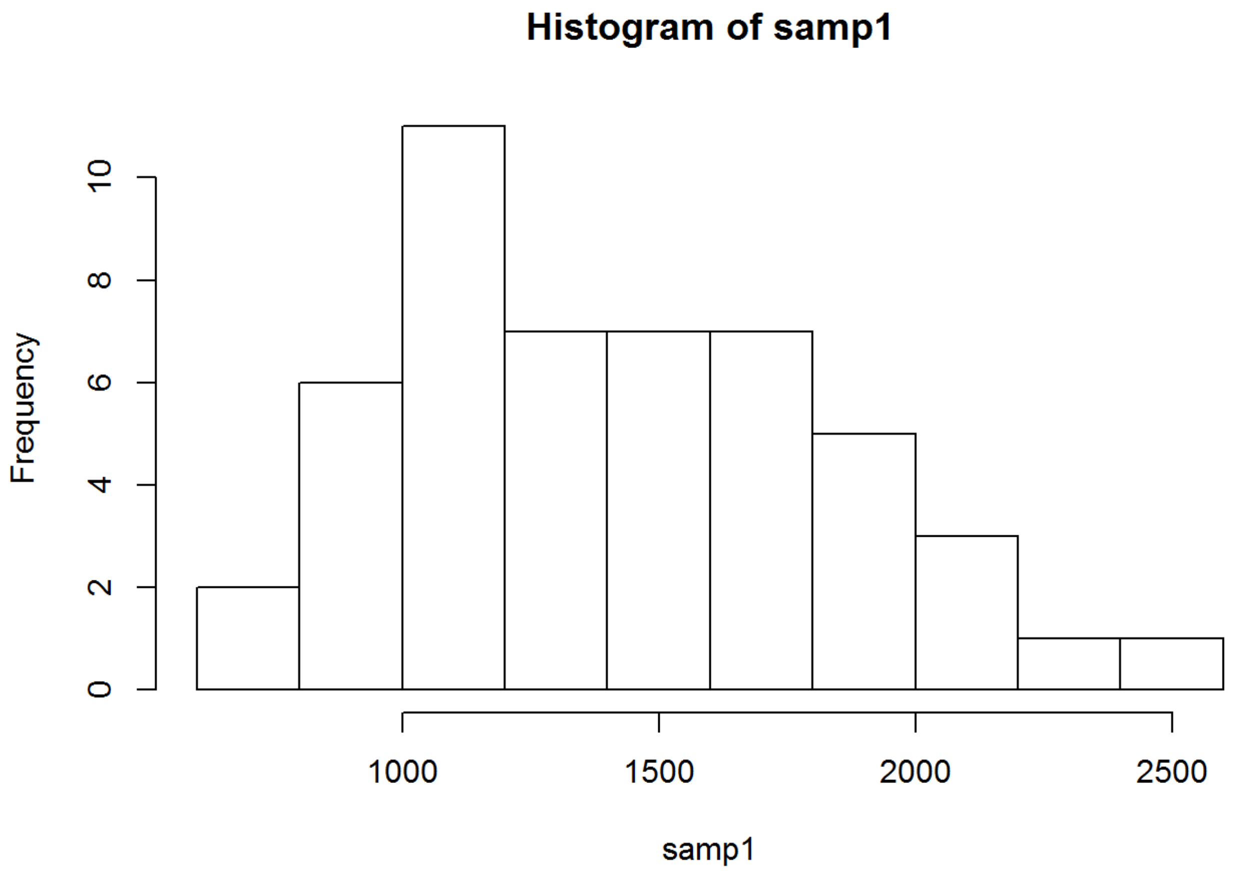


Exercise 1: Describe this population distribution. According to the histogram, the distribution appear to be right skewed with peak at between 1000 to 1500 square feet. The range of the area is between 0 and 4000 square feet.

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

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

```
hist(samp1)
```



The distribution is quite similar to the population distribution. It has only one peak and it is right skewed.

```
mean(samp1)
```

```
## [1] 1433.48
```

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

```
## [1] 1496.48
```

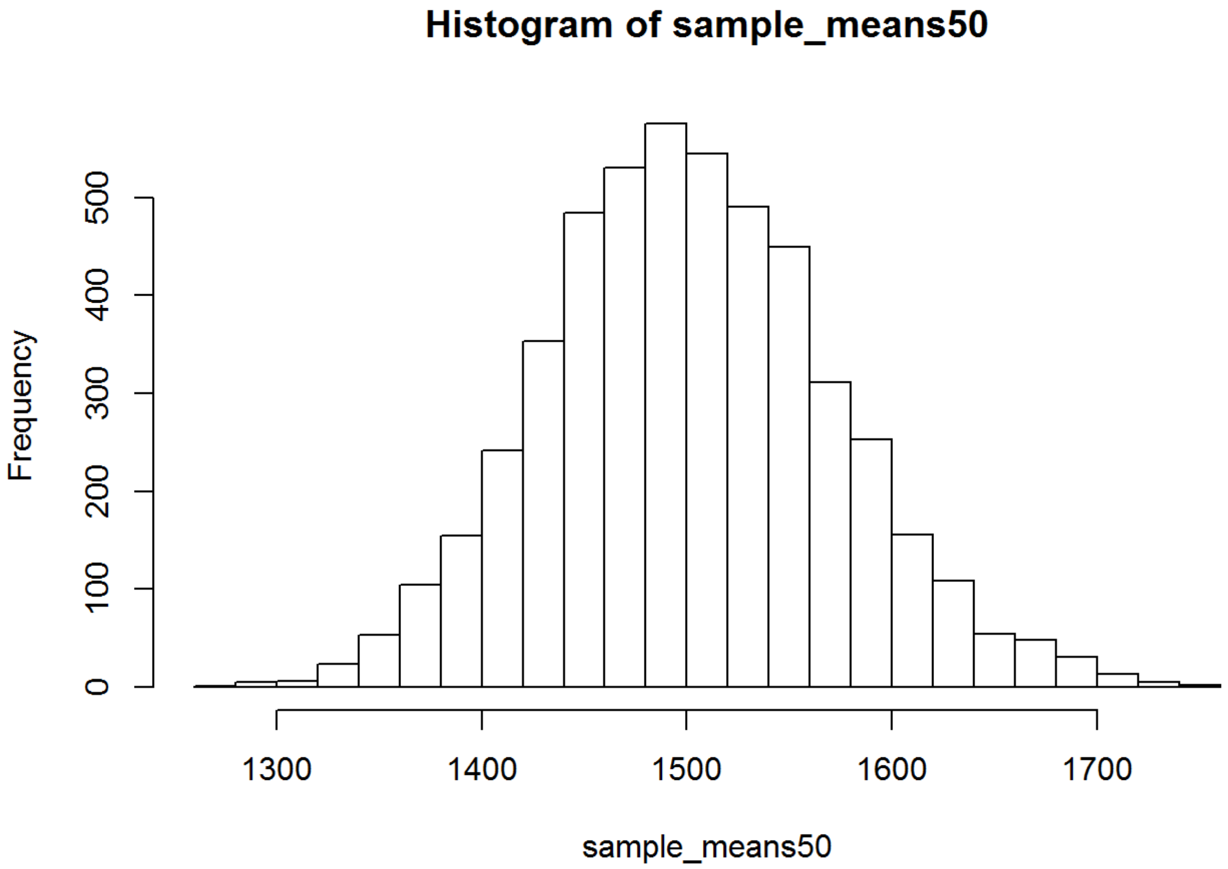
Exercise 3: Take a second sample, also of size 50, and call it samp2. How does the mean of samp2 compare with the mean of samp1? Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population mean?

samp1 has mean of 1532.68, samp2 has mean of 1486.98 Both means are closed to the population mean which is 1499.69 square feet. sam2 is a little closer than samp1.The sample with size of 1000 will have more accurate estimate of the population mean. Because the higher the sample size, the smaller the standard error of the mean (SE= s/sqrt(n)), so more accurate the estimate of the population mean.

```
sample_means50 <- rep(NA, 5000)

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

hist(sample_means50, breaks = 25)
```



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? There are 5000 elements in sample_means50. The distribution is unimodal, symmetric, with its center at around 1500 square feet. We can conclude that the distribution of the sample mean is normal. If we collect more sample means, the center of the graph will be getting closer and closer to 1499.69. It will become more and more symmetric and normal.

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(0, 100)

for (i in 1:100) {
  samp <- sample(area, 50)
  sample_means_small[i] <- mean(samp)
}
sample_means_small
```

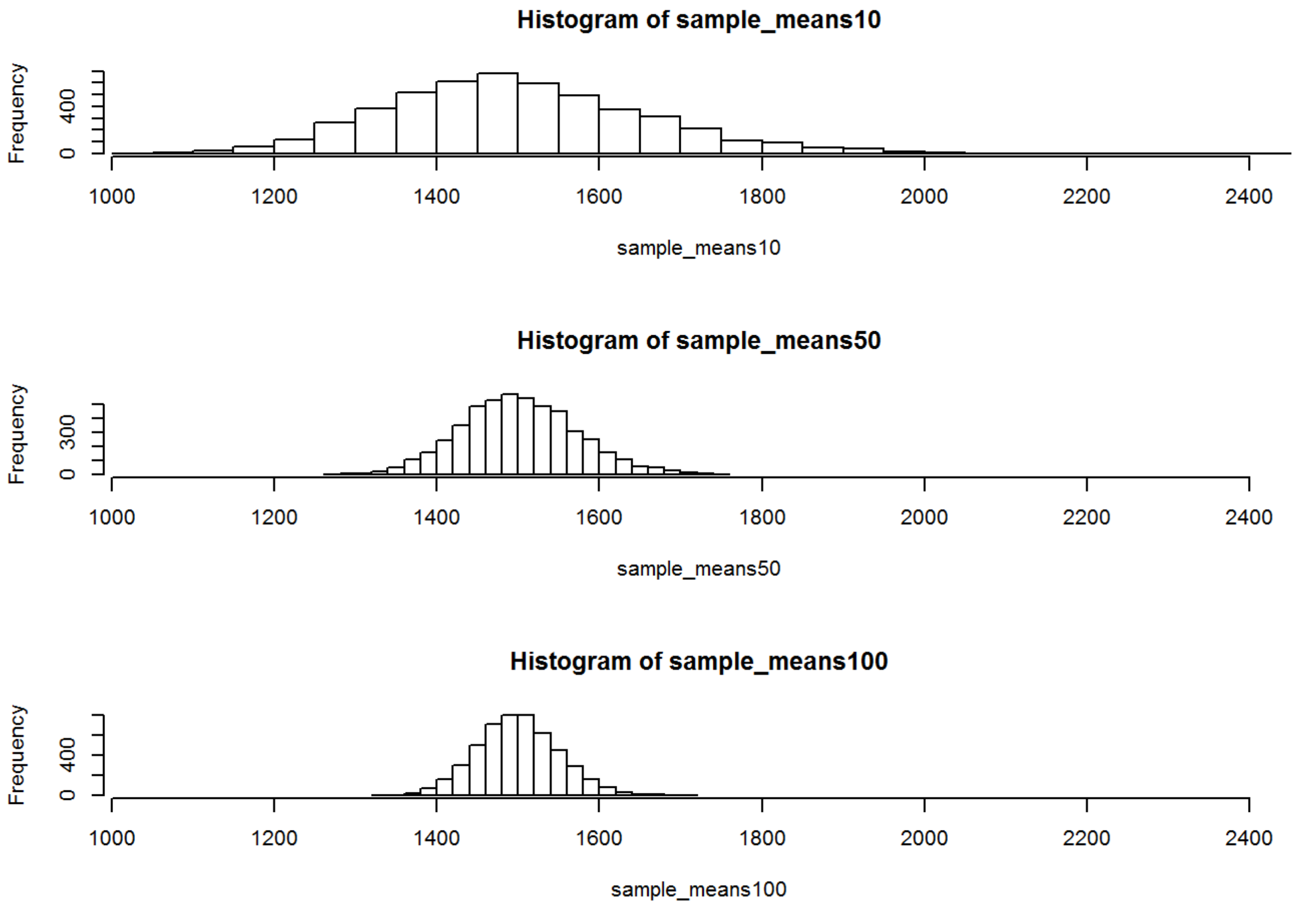
```
##   [1] 1461.02 1453.28 1521.90 1411.90 1629.82 1529.76 1429.68 1496.26
##   [9] 1506.16 1472.96 1616.40 1416.44 1650.48 1593.36 1346.80 1556.44
##  [17] 1488.82 1452.22 1607.82 1457.40 1439.16 1521.98 1443.86 1427.46
##  [25] 1448.82 1365.48 1458.80 1507.66 1498.78 1554.48 1482.56 1453.06
##  [33] 1488.00 1588.02 1574.98 1456.46 1556.30 1455.80 1462.46 1529.76
##  [41] 1542.92 1435.62 1505.18 1509.54 1503.68 1567.98 1515.18 1608.22
##  [49] 1581.80 1577.60 1499.80 1517.06 1560.72 1529.48 1461.54 1504.12
##  [57] 1489.80 1672.08 1508.72 1568.56 1500.54 1612.78 1435.20 1599.34
##  [65] 1613.52 1487.32 1434.90 1448.60 1512.02 1557.36 1497.92 1471.76
##  [73] 1446.98 1504.98 1459.04 1490.94 1578.14 1442.12 1539.06 1574.98
##  [81] 1506.00 1425.20 1505.28 1522.86 1622.36 1552.46 1553.76 1446.82
##  [89] 1501.14 1458.46 1490.76 1626.88 1486.90 1436.60 1473.34 1727.28
##  [97] 1465.28 1419.92 1511.24 1550.32
```

There are total 100 elements in variable sample_means_small. Each element represent the average area of 50 houses in the random sample.

```
sample_means10 <- rep(NA, 5000)
sample_means100 <- rep(NA, 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 6: When the sample size is larger, what happens to the center? What about the spread?

The center will be getting closer to the true population mean. The spread of the distribution will be getting more and more narrow because the standard deviation of the sample mean will be getting smaller and smaller.

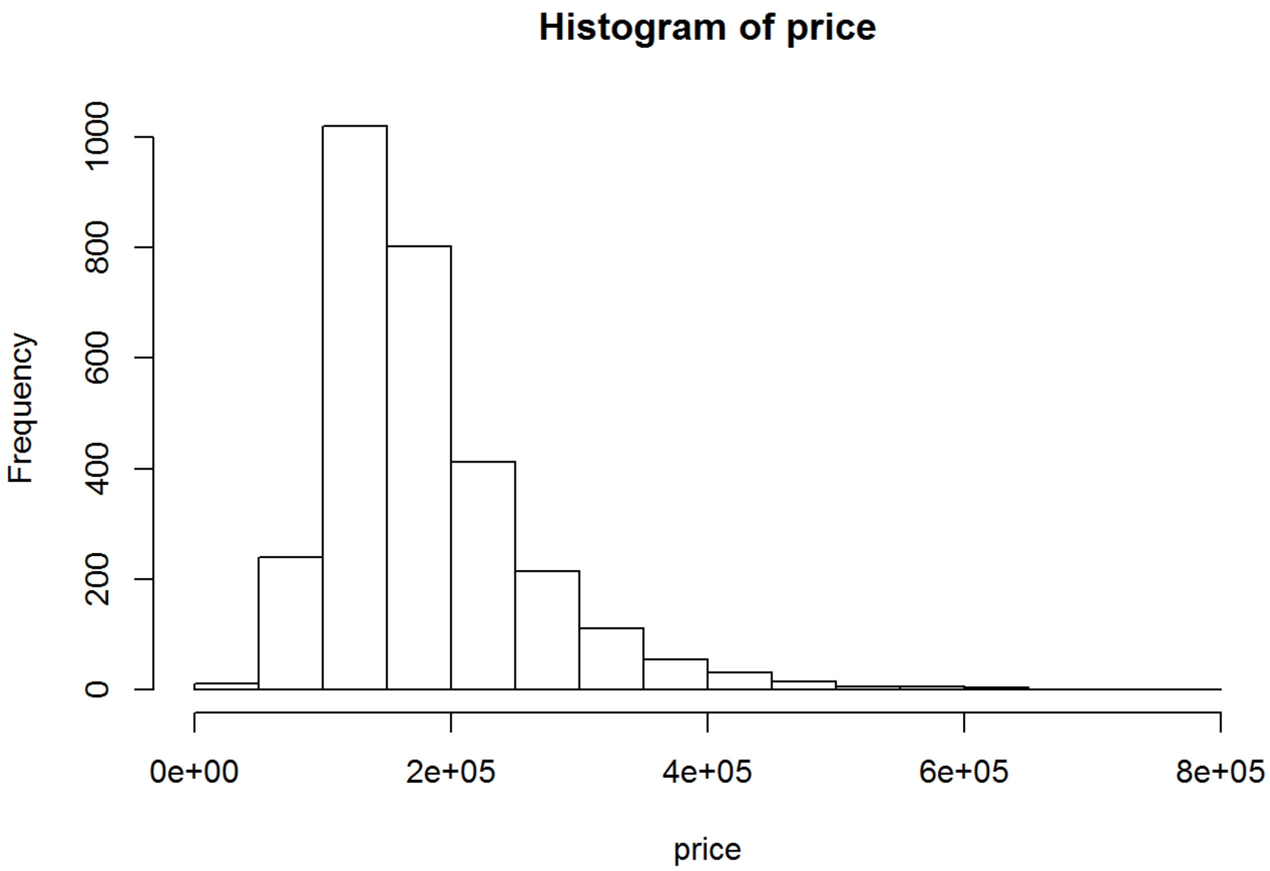
On your own So far, we have only focused on estimating the mean living area in homes in Ames. Now you'll try to estimate the mean home price.

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

```
summary(price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   12790  129500  160000  180800  213500  755000
```

```
hist(price)
```



```
samp1 <- sample(price, 50)
mean(samp1)
```

```
## [1] 172857.7
```

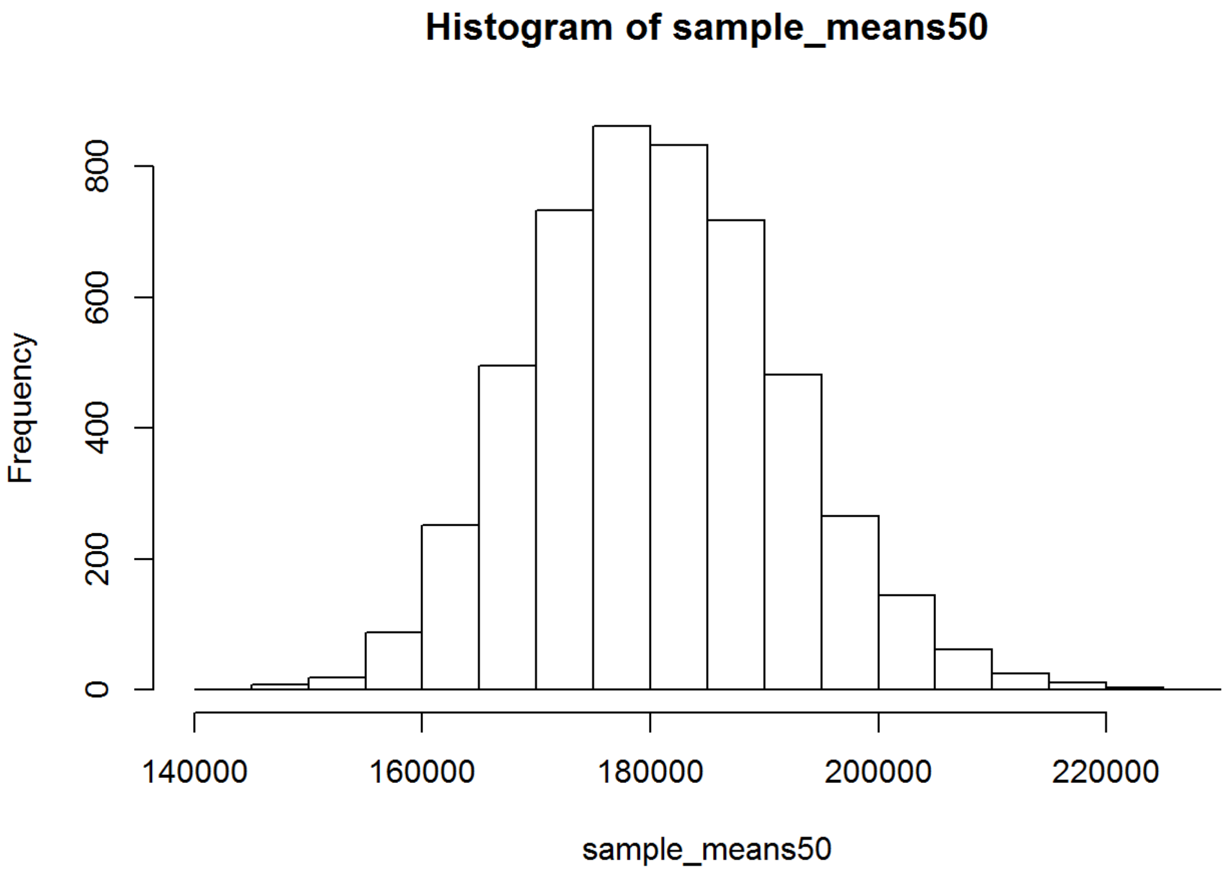
The best point estimate of the population mean is the sample mean which is 185064.4 dollars.

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.

```
sample_means50 <- rep(NA, 5000)

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

hist(sample_means50)
```



```
mean(sample_means50)
```

```
## [1] 180714.1
```

```
mean(price)
```

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

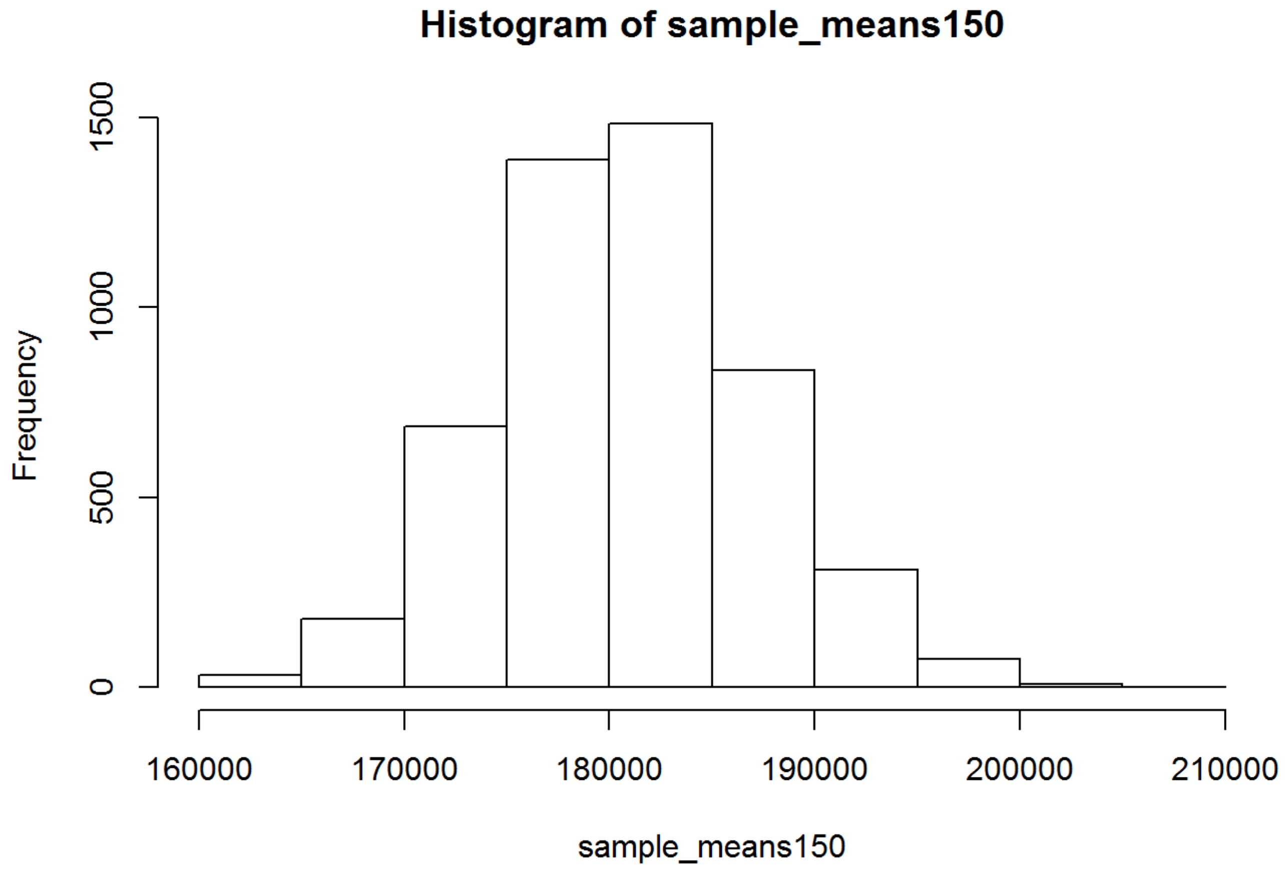
Based on the graph, the sampling distribution is unimodal, symmetric and resembles normal distribution. The center of the graph is about 180000 dollars. When I calculated the sample mean home price of the population is 180787.3 dollars. The calculation of the true population mean is actually 180796.1, which is closed to my estimate.

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?

```
sample_means150 <- rep(NA, 5000)

for(i in 1:5000){
   samp <- sample(price, 150)
   sample_means150[i] <- mean(samp)
   }

hist(sample_means150)
```



```
mean(sample_means150)
```

```
## [1] 180784.9
```

```
mean(price)
```

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

This sampling distribution of sample size 150 has lower variability. Its curve appears to be more narrow compare to my previous sampling distribution. Also in addition, its sample mean is 180810.5, which is also closer to my population mean (180796.1)

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? Sampling distributiom from 3 has smaller spread. we would prefer a distribution with small spread, since it will be closer to the true population mean, also less standard deviation every time we pick a sample space.