Lab 3A: Foundataions for inference - Sampling Disributions

Charles Brown

Sunday, March 29, 2015

Contents

0.0.1 Load Data

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

0.0.2 Exploratory Stats

```
sort(names(ames))
```

```
##  [1] "Alley"           "Bedroom.AbvGr"   "Bldg.Type"
##  [4] "Bsmt.Cond"       "Bsmt.Exposure"   "Bsmt.Full.Bath"
##  [7] "Bsmt.Half.Bath"  "Bsmt.Qual"       "Bsmt.Unf.SF"
## [10] "BsmtFin.SF.1"    "BsmtFin.SF.2"    "BsmtFin.Type.1"
## [13] "BsmtFin.Type.2"  "Central.Air"     "Condition.1"
## [16] "Condition.2"     "Electrical"      "Enclosed.Porch"
## [19] "Exter.Cond"      "Exter.Qual"      "Exterior.1st"
## [22] "Exterior.2nd"    "Fence"           "Fireplace.Qu"
## [25] "Fireplaces"      "Foundation"      "Full.Bath"
## [28] "Functional"      "Garage.Area"     "Garage.Cars"
## [31] "Garage.Cond"     "Garage.Finish"   "Garage.Qual"
## [34] "Garage.Type"     "Garage.Yr.Blt"   "Gr.Liv.Area"
## [37] "Half.Bath"       "Heating"         "Heating.QC"
## [40] "House.Style"     "Kitchen.AbvGr"   "Kitchen.Qual"
## [43] "Land.Contour"    "Land.Slope"      "Lot.Area"
## [46] "Lot.Config"      "Lot.Frontage"    "Lot.Shape"
## [49] "Low.Qual.Fin.SF" "Mas.Vnr.Area"    "Mas.Vnr.Type"
## [52] "Misc.Feature"    "Misc.Val"        "Mo.Sold"
## [55] "MS.SubClass"     "MS.Zoning"       "Neighborhood"
## [58] "Open.Porch.SF"   "Order"           "Overall.Cond"
## [61] "Overall.Qual"    "Paved.Drive"     "PID"
## [64] "Pool.Area"       "Pool.QC"         "Roof.Matl"
## [67] "Roof.Style"      "Sale.Condition"  "Sale.Type"
## [70] "SalePrice"       "Screen.Porch"    "Street"
## [73] "Total.Bsmt.SF"   "TotRms.AbvGrd"   "Utilities"
## [76] "Wood.Deck.SF"    "X1st.Flr.SF"     "X2nd.Flr.SF"
## [79] "X3Ssn.Porch"     "Year.Built"      "Year.Remod.Add"
## [82] "Yr.Sold"
```

```r
#
area <- ames$Gr.Liv.Area
(xsum <-summary(area))
```

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

```r
hist(x = area, xlab = "Area in Sq. Feet", breaks = 100)
text(4000, 120, paste("Mean =", round(mean(area), 1), "\n Median =",
         round(median(area), 1), "\n Std.Dev =", round(sd(area), 1)))
```

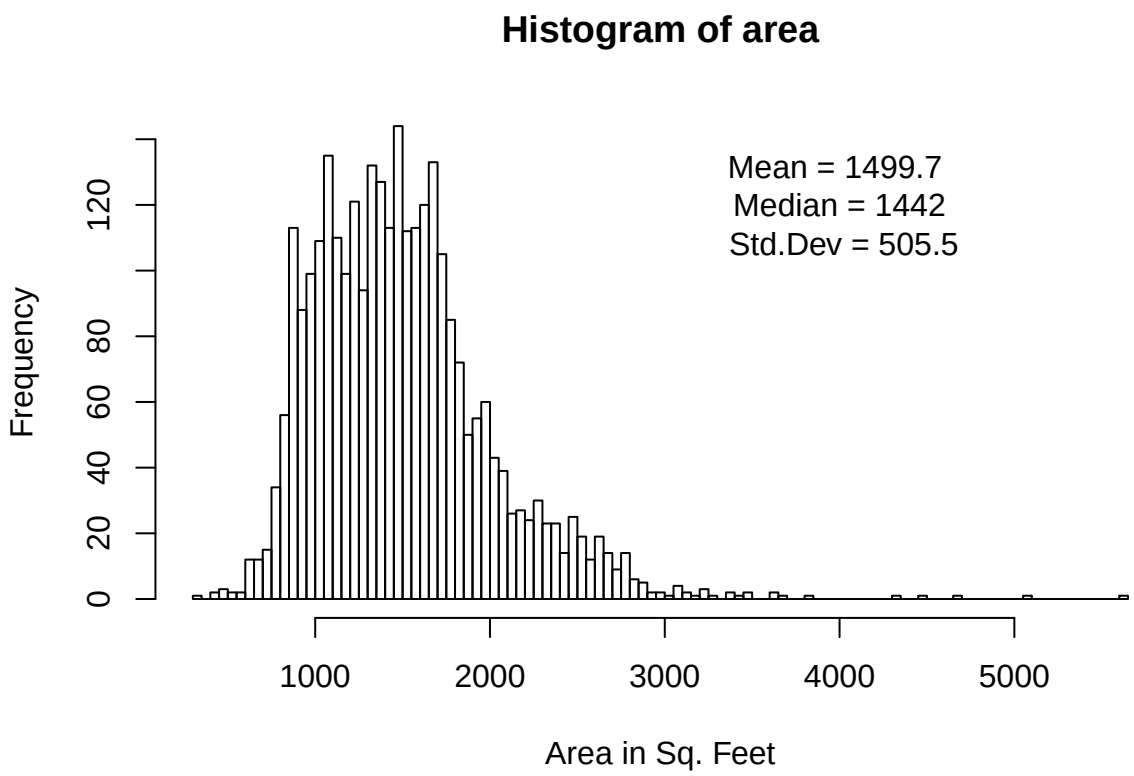

Figure 1:

0.0.2.1 Question 1: Which of the following is false?

A) The distribution of areas of houses in Ames is unimodal and right-skewed.
B) **50% of houses in Ames are smaller than 1,500 square feet.**
C) The middle 50% of the houses range between approximately 1,130 square feet and 1,740 square feet.
D) The IQR is approximately 610 square feet.
E) The smallest house is 334 square feet and the largest is 5,642 square feet.

- Half of the houses in Ames are less than the median of 1442 square feet.

Question 1 Answer: B)

0.0.3 The Unknown Sampling Distribution.

```
samp0 <-sample(area,50)
samp1 <-sample(area,50)
#
mean(samp1)
```

```
## [1] 1519.24
```

```
summary(samp1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     672    1082    1437    1519    1847    2798
```

```
#
samp_50 <-sample(area,50)
samp_100 <-sample(area,100)
samp_1000 <-sample(area, 1000)
```

0.0.3.1 Question 2: 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?

A) Sample size of 50
B) Sample size of 100
C) **Sample size of 1000**

- Mean of Entire Population of : 1499.6904437.
- Mean of Sample size 50: 1421.76.
- Mean of Sample size 100: 1480.64.
- Mean of sample size 1000: 1518.348.

Question 2 Answer: The larger the sample size, the better the estimate of the mean.

The distribution of sample means, called the sampling distribution, can help us understand this variability. In this lab, because we have access to the population, we can build up the sampling distribution for the sample mean by repeating the above steps many times. Here we will generate 5000 samples and compute the sample mean of each.

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

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

(xsum <-summary(sample_means50))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1265    1451    1499    1501    1548    1771
```

```
hist(sample_means50, breaks = 50)
text(1725, 250, paste("Mean =", round(mean(sample_means50), 1), "\n Median =",
        round(median(sample_means50), 1), "\n Std.Dev =", round(sd(sample_means50), 1)))
```

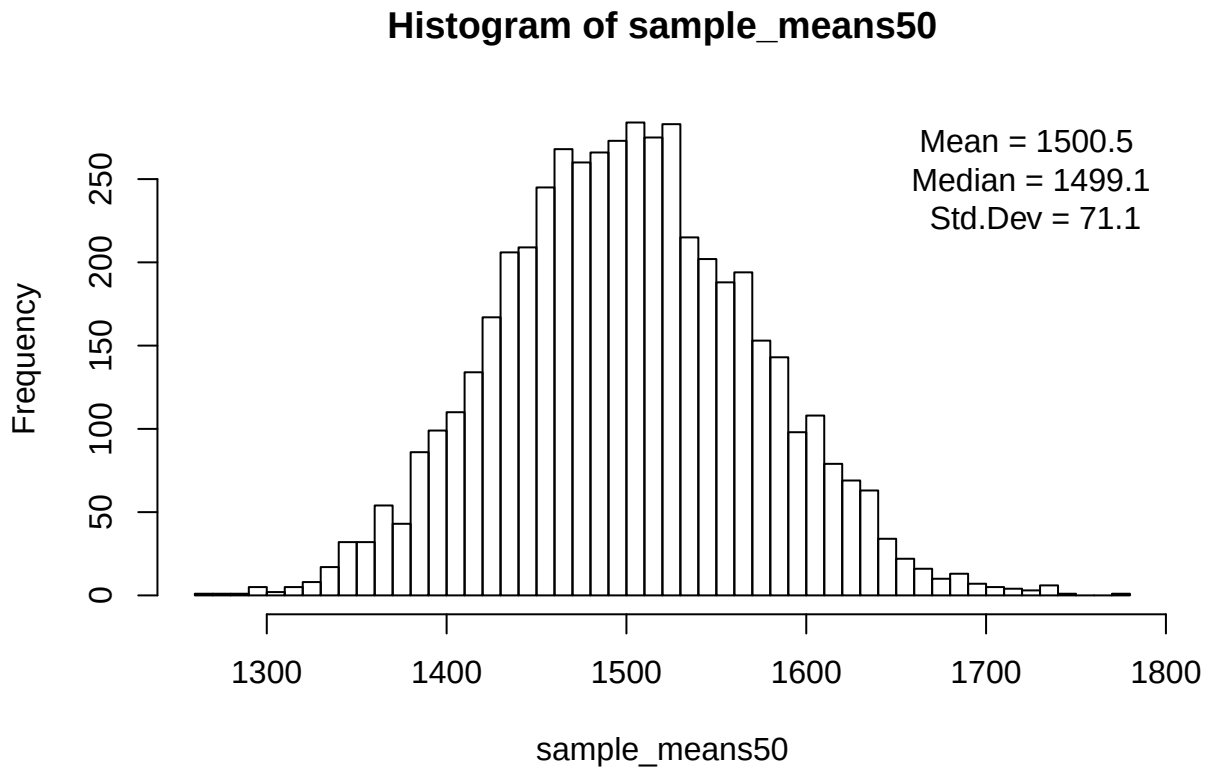

Figure 2:

```
summary(sample_means50)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1265    1451    1499    1501    1548    1771
```

0.0.3.2 Exercise: Describe the sampling distribution (the distribution of the sample means that you just created), and be sure to specifically note its center. Excercise Answer: The sampling distribution looks a like a normal distribution with a mean of 1500.506272.

```
sample_means_small <-rep(NA, 100)
```

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

}

sample_means_small
```

0.0.3.3 Exercise: To make sure you understand what you've done in this loop, try running a smaller version. Initialize a vector of 100 NAs 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. Print the output to your screen (type sample_means_small into the console and press enter).

```
##   [1] 1440.52 1456.38 1546.40 1453.22 1440.46 1495.98 1469.60 1407.00
##   [9] 1532.84 1558.88 1552.46 1614.68 1468.86 1572.72 1588.06 1349.74
##  [17] 1550.36 1483.44 1451.48 1458.22 1549.62 1670.06 1631.80 1438.94
##  [25] 1588.66 1522.04 1508.46 1509.82 1496.66 1539.38 1514.88 1437.16
##  [33] 1377.64 1456.62 1424.04 1525.96 1508.68 1578.52 1446.42 1425.92
##  [41] 1470.88 1480.84 1508.82 1511.80 1494.22 1466.74 1410.76 1463.52
##  [49] 1441.52 1555.68 1487.50 1422.34 1636.46 1458.50 1479.40 1376.98
##  [57] 1477.62 1458.36 1556.12 1470.72 1553.48 1612.04 1422.36 1572.46
##  [65] 1527.26 1478.82 1492.30 1391.20 1503.06 1629.36 1468.34 1618.38
##  [73] 1437.18 1546.74 1427.26 1346.76 1428.58 1471.82 1532.70 1675.44
##  [81] 1553.94 1533.92 1469.98 1420.28 1441.96 1368.98 1445.30 1358.36
##  [89] 1570.08 1539.12 1541.44 1432.80 1611.60 1467.78 1559.06 1600.00
##  [97] 1385.70 1620.50 1418.54 1482.94
```

```
hist(sample_means_small, breaks = 25)
text(1650, 7, paste("Mean =", round(mean(sample_means_small), 1), "\n Median =",
         round(median(sample_means_small), 1), "\n Std.Dev =", round(sd(sample_means_small), 1)))
```

0.0.3.4 Question 3: How many elements are there in this object called sample_means_small?

A) 0
B) 30
C) 50
D) **100**
E) 5,000

Question 3 Answer: `sample_means_small` contains 100 elements.

0.0.3.5 Question 4: Which of the following is true about the elements in the sampling distributions you created?

A) **Each element represents a mean square footage from a simple random sample of 50 houses.**
B) Each element represents the square footage of a house.
C) Each element represents the true population mean of square footage of houses.

Question 4 Answer: Each element of this sampling distribution vector is a mean of a random sample (size 50) from the population.

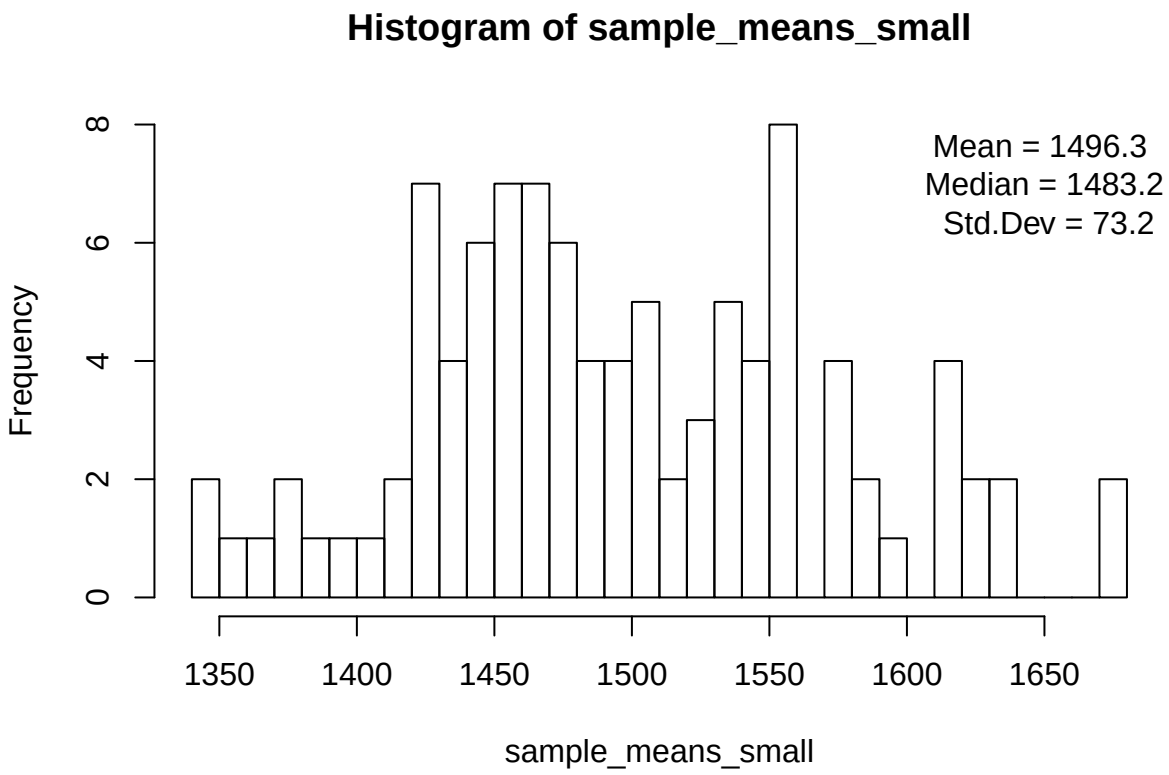
Histogram of sample_means_small
Mean = 1496.3
Median = 1483.2
Std.Dev = 73.2
Frequency
0
2
4
6
8
1350
1400
1450
1500
1550
1600
1650
sample_means_small

Figure 3:

0.0.4 Sample size and the sampling distribution

To get a sense of the effect that sample size has on our sampling distribution, let's build up two more sampling distributions: one based on a sample size of 10 and another based on a sample size of 100.

```
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)
}
```

To see the effect that different sample sizes have on the sampling distribution, plot the three distributions on top of one another.

```
par(mfrow = c(3, 1))
xlimits = range(sample_means10)
hist(sample_means10, breaks = 20, xlim = xlimits)
text(1850, 500, paste("Std.Dev =", round(sd(sample_means10), 1)))
hist(sample_means50, breaks = 20, xlim = xlimits)
text(1850, 500, paste("Std.Dev =", round(sd(sample_means50), 1)))
hist(sample_means100, breaks = 20, xlim = xlimits)
text(1850, 500, paste("Std.Dev =", round(sd(sample_means100), 1)))
```

0.0.4.1 Question 5: It makes intuitive sense that as the sample size increases, the center of the sampling distribution becomes a more reliable estimate for the true population mean. Also as the sample size increases, the variability of the sampling distribution _____.

A) **decreases**
B) increases
C) stays the same

Answer Question 5: The panel plot shows that the variability decreases as the sample size increases.

0.0.5 Now you'll try to estimate the mean home price.

```
price <- ames$SalePrice
samp <-sample(price,50)
(xsum <-summary(samp))
```

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

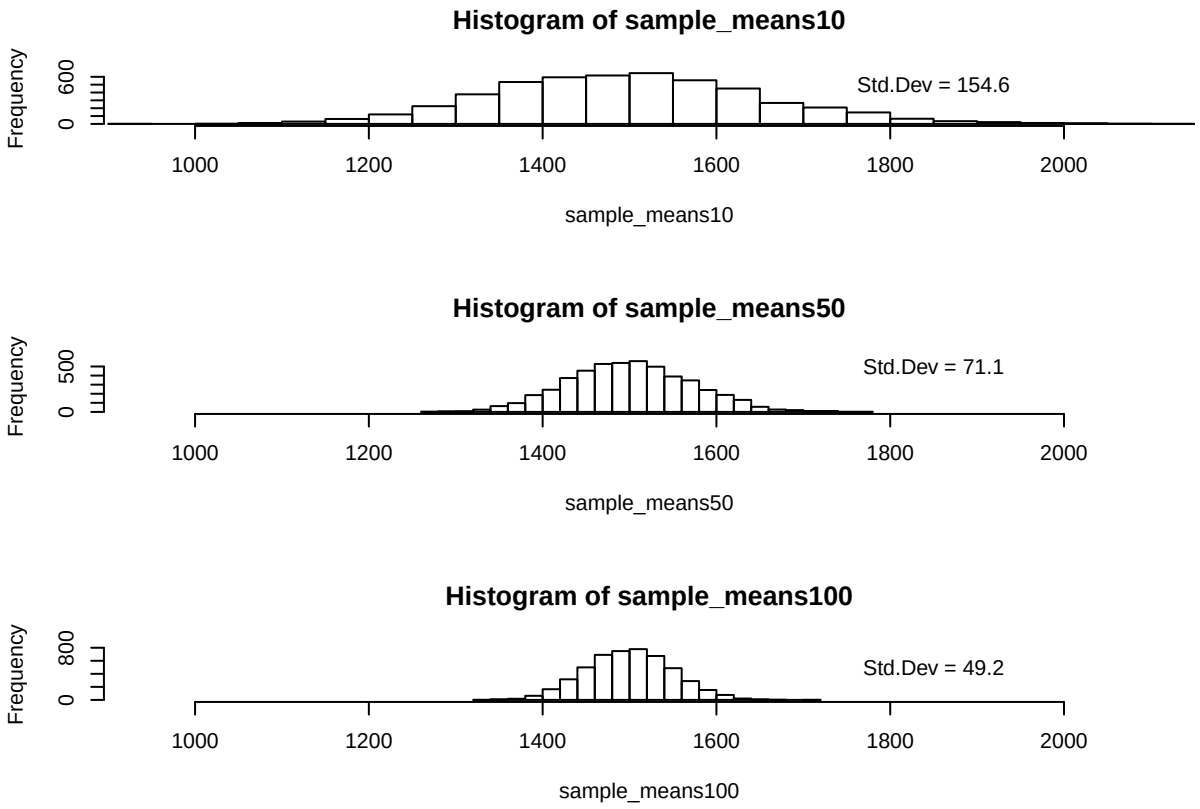
Histogram of sample_means10
Frequency
0
600
Std.Dev = 154.6
1000
1200
1400
1600
1800
2000
sample_means10
Histogram of sample_means50
Frequency
0
500
Std.Dev = 71.1
1000
1200
1400
1600
1800
2000
sample_means50
Histogram of sample_means100
Frequency
0
800
Std.Dev = 49.2
1000
1200
1400
1600
1800
2000
sample_means100

Figure 4:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   78000  136600  165800  188200  229600  378500
```

```
hist(samp, xlab = "Price", breaks = 50)
```

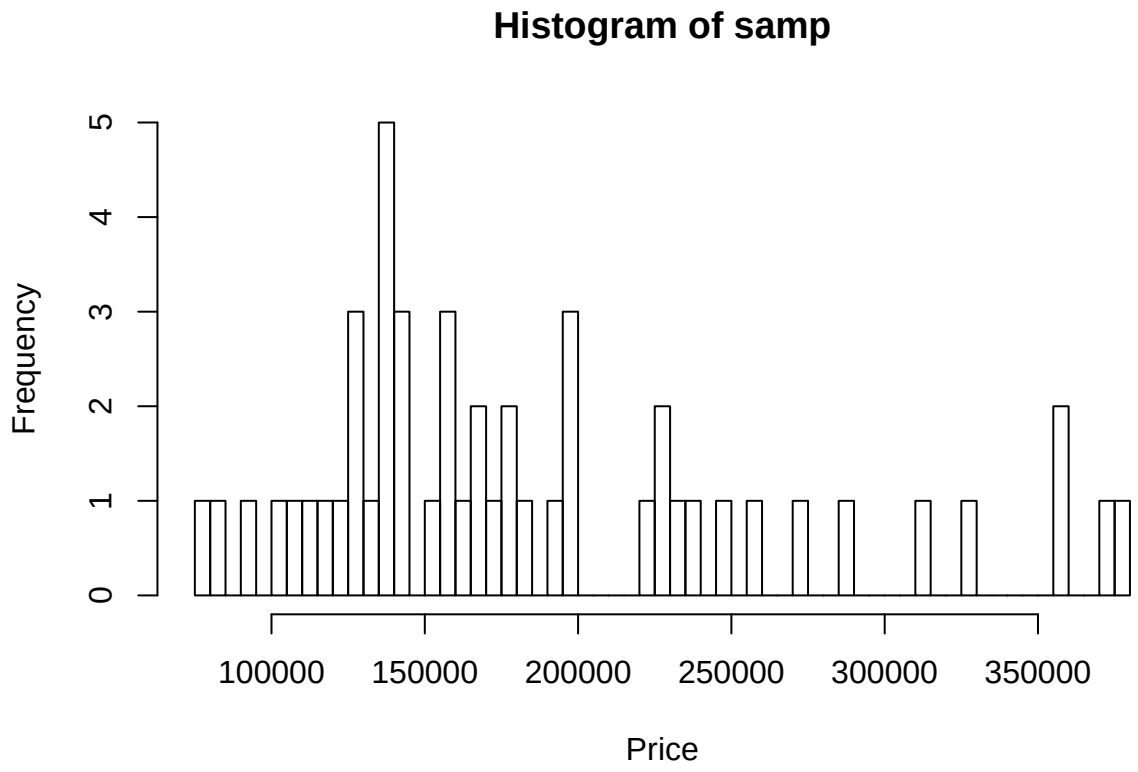

Figure 5:

Exercise Answer: The point estimate of the mean is 188.2K.

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

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

hist(sample_means50, breaks = 50)
```

```
(xsum50 <-summary(sample_means50))
```

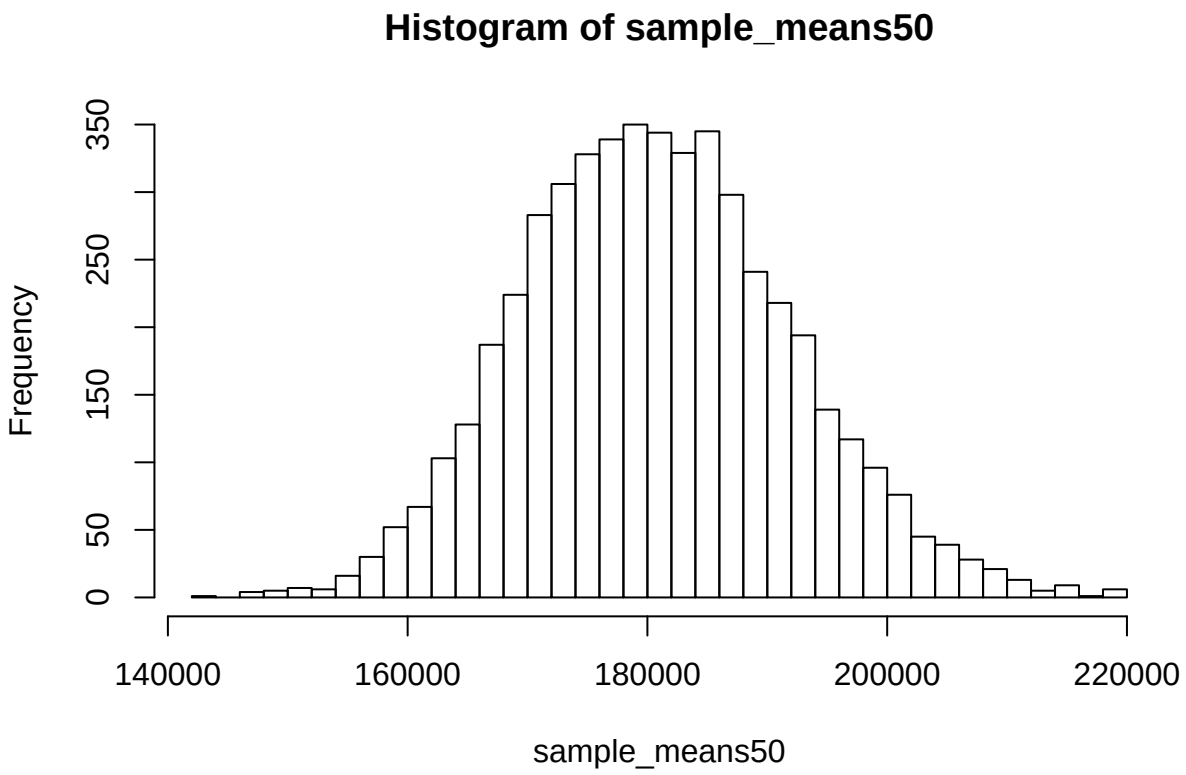
Histogram of sample_means50
Frequency
0
50
150
250
350
140000
160000
180000
200000
220000
sample_means50

Figure 6:

0.0.5.2 Exercise: Since you have access to the population, simulate the sampling distribution for $\bar{x}$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?

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  143300  172900  180400  180800  188000  219500
```

Exercise Answer: The shape of the price sampling distribution is normal because the median home price of 180.4K is approximately the same as the mean home price of 180.8K.

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

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

hist(sample_means150, breaks = 50)
```

```
(xsum150 <-summary(sample_means150))
```

0.0.5.3 Exercise: 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?

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  160300  176500  180600  180800  185100  205400
```

```
par(mfrow = c(2,1))
xlimits50 <-range(sample_means50)
hist(sample_means50, breaks = 50, xlim = xlimits50)
text(210000, 200, paste("Std.Dev =", round(sd(sample_means50), 1)))
hist(sample_means150, breaks = 50, xlim = xlimits50)
text(210000, 200, paste("Std.Dev =", round(sd(sample_means150), 1)))
```

- Sample Size 50 Summary:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  143300  172900  180400  180800  188000  219500
```

- Sample Size 150 Summary:

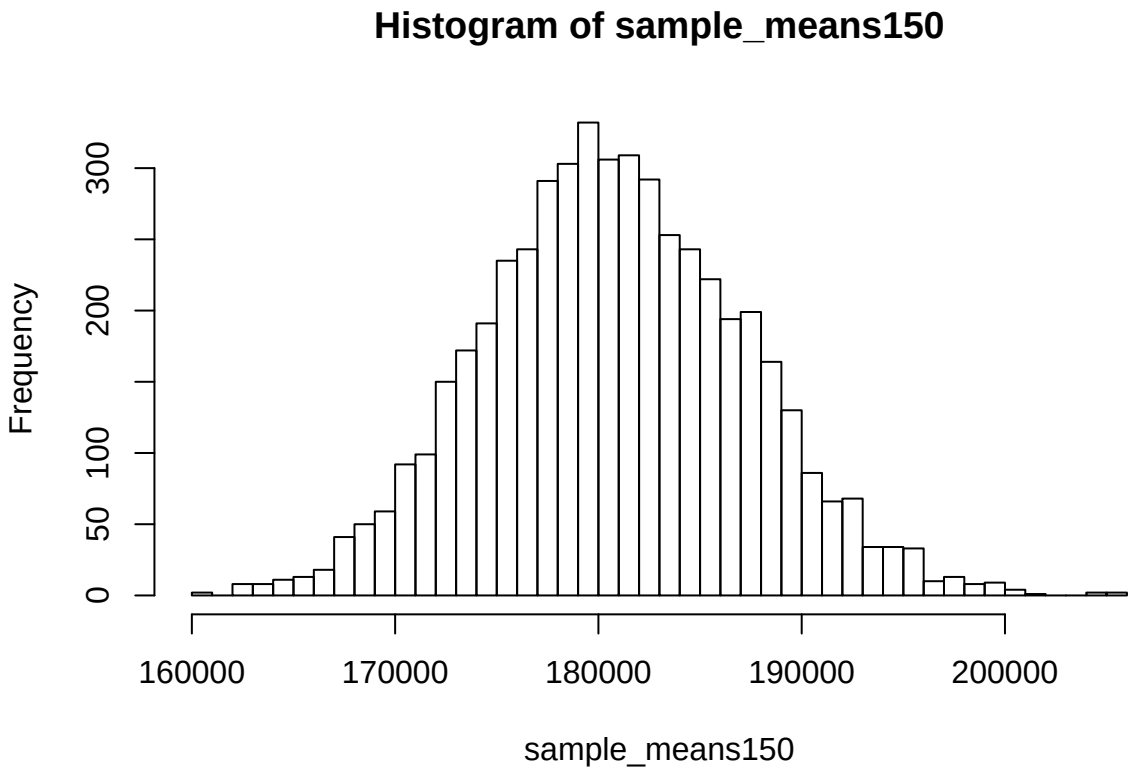
Histogram of sample_means150
Frequency
0
50
100
200
300
160000
170000
180000
190000
200000
sample_means150

Figure 7:

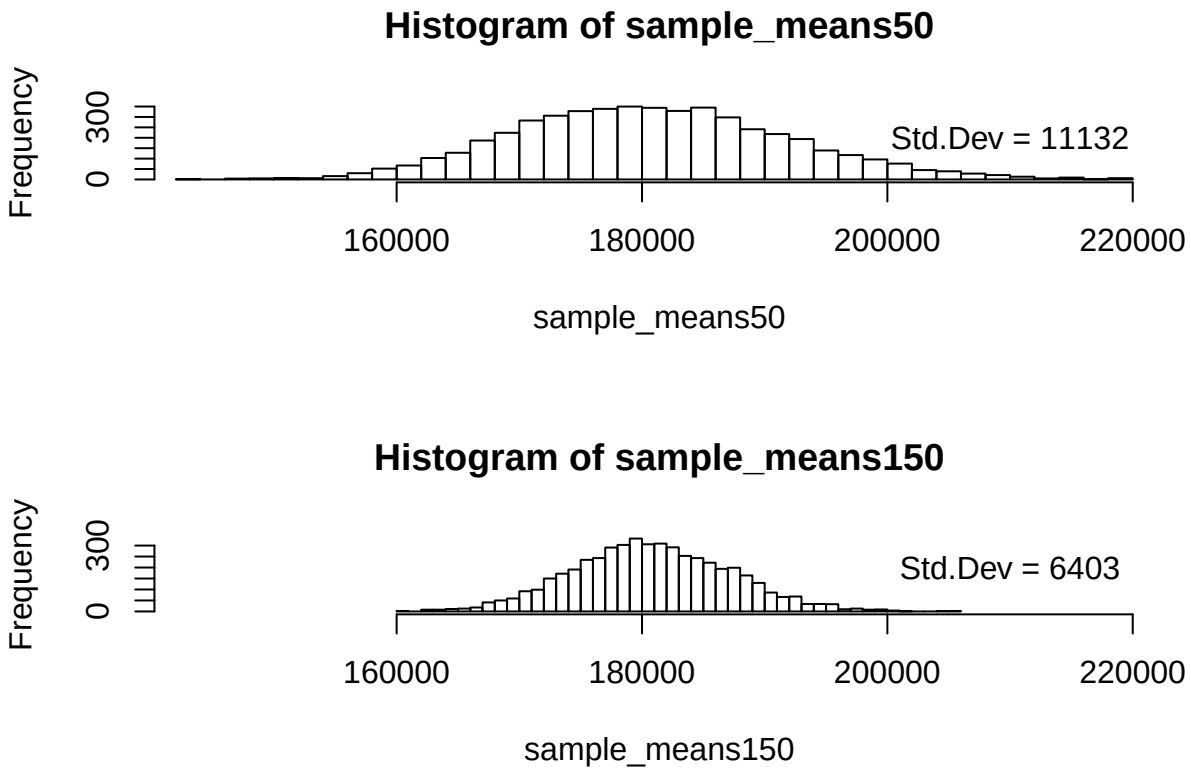
Histogram of sample_means50
Frequency
0
300
Std.Dev = 11132
160000
180000
200000
220000
sample_means50
Histogram of sample_means150
Frequency
0
300
Std.Dev = 6403
160000
180000
200000
220000
sample_means150

Figure 8:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  160300  176500  180600  180800  185100  205400
```

Exercise Answer: Both sampling distributions have a near normal shape. The estimated mean price of homes in Ames is $184500.

0.0.5.4 Question 6: Which of the following is false?

A) **The variability of the sampling distribution with the smaller sample size (sample_means50) is smaller than the variability of the sampling distribution with the larger sample size (sample_means150).**
B) The means for the two sampling distribtuions are roughly similar.
C) Both sampling distributions are symmetric.

Question 6 Answer: A is false. Sample sizes of 150 typically have a smaller sampling distribution spread (45100) than sample sizes of 50 (300500).
