Hao-4a

Bruce Hao

September 16, 2016

```r
#setwd("C:/Users/bhao/Google Drive/CUNY/git/DATA606/Lab4a")
library(IS606)
library(dplyr)
library(ggplot2)
load("more/ames.RData")
area <- ames$Gr.Liv.Area
price <- ames$SalePrice
samp1 = sample(area, 50)
samp2 = sample(area, 50)
```

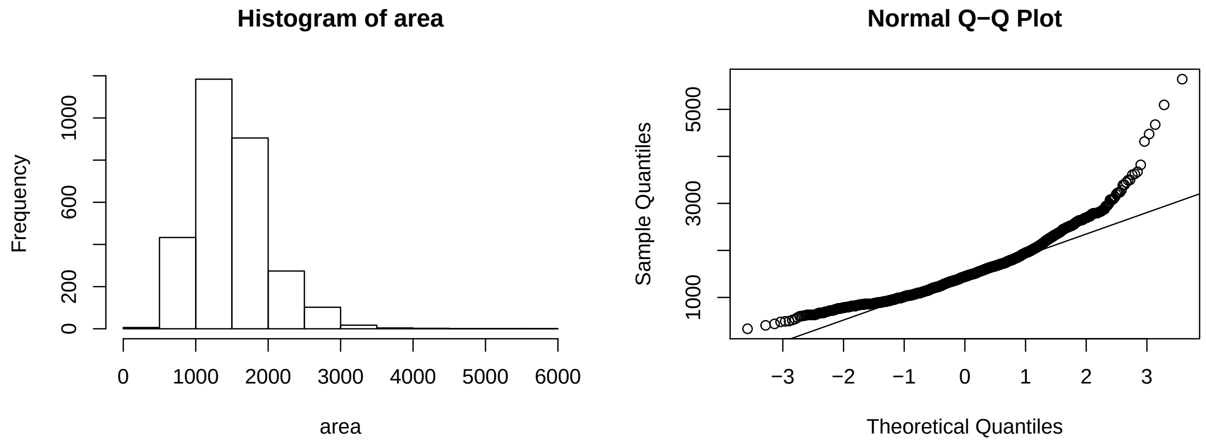
Exercise 1: Describe this population distribution.

Based on the charts below, the area data appear right skewed.

```r
summary(area)
```

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

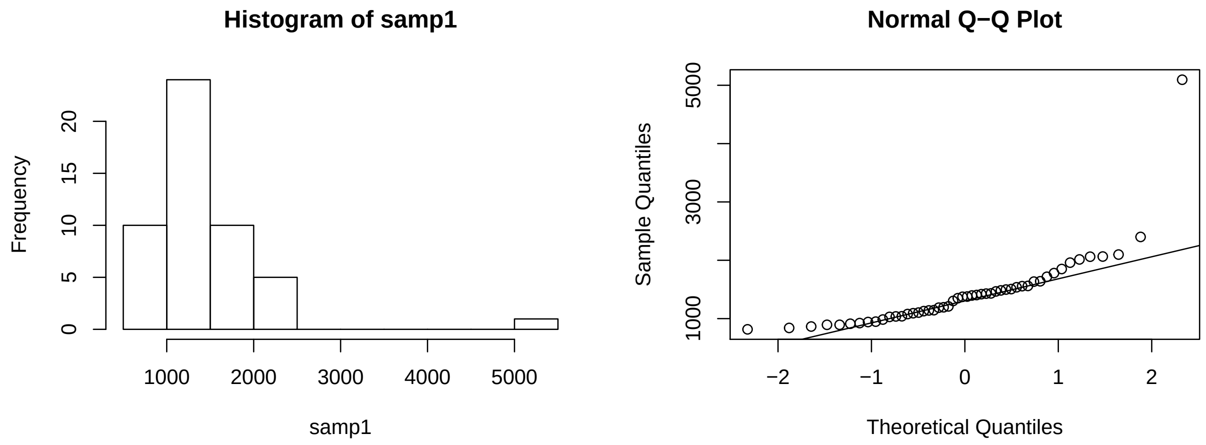
```r
par(mfrow = c(1, 2))
hist(area)
qqnorm(area)
qqline(area)
```



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

The sample data appears to share the same distribution as the population data.

```r
par(mfrow = c(1, 2))
hist(samp1)
qqnorm(samp1)
qqline(samp1)
```



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?

The mean of samp2 is a bit greater than the population mean, while the mean of samp1 is a bit below. The larger the sample size, the more closely the mean of the sample should be to the mean of the population.

```r
mean(area)
```

```
## [1] 1499.69
```

```r
mean(samp1)
```

```
## [1] 1435.96
```

```r
mean(samp2)
```

```
## [1] 1537.22
```

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 normally distributed and centered at 1500, which is very close to the population mean. As more sample means are collected, 1) the mean of sample means would converge to the population mean, 2) the standard deviation of sample means would decrease and 3) the distribution of sample means would become more and more normally distributed.

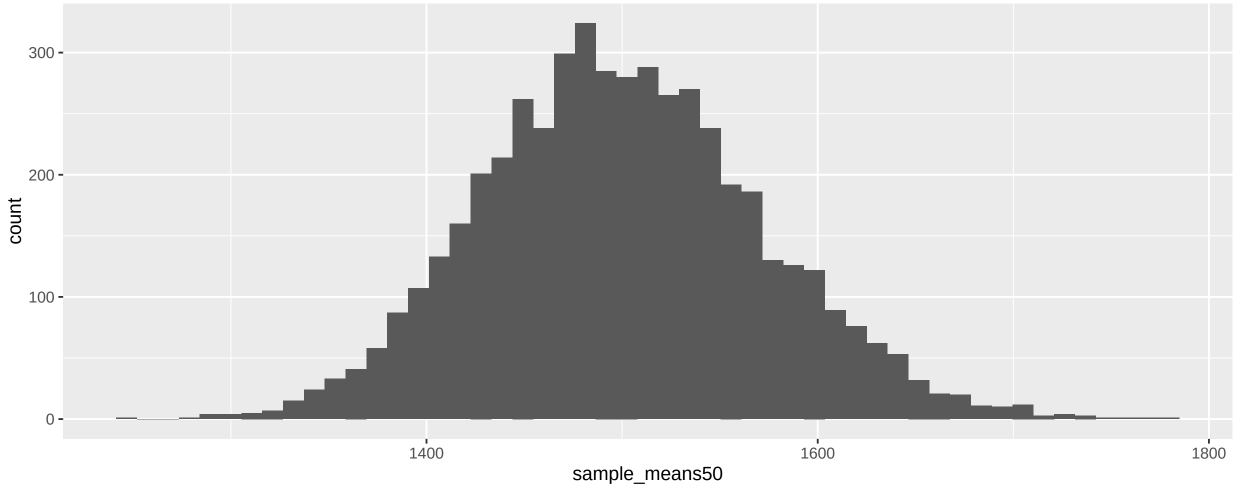
```
sample_means50 <- rep(NA, 5000)

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

mean(sample_means50)
```

```
## [1] 1499.953
```

```
#data.frame(sample_means50) %>% ggplot(aes(sample_means50)) + geom_histogram(binwidth = 10)
qplot(sample_means50, geom = 'histogram', binwidth = (max(sample_means50)-min(sample_means50))/50)
```



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?

There are 100 elements in sample_means_small, each representing the average area of a random sample of 50 houses from the population.

```
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] 1653.74 1476.24 1489.10 1430.46 1496.14 1566.64 1504.18 1605.32
##   [9] 1501.98 1503.84 1552.84 1562.56 1553.58 1431.50 1448.82 1570.40
##  [17] 1438.20 1515.74 1480.26 1481.64 1492.14 1528.92 1528.48 1548.36
##  [25] 1482.30 1480.86 1484.50 1440.24 1518.20 1571.72 1443.68 1558.18
##  [33] 1510.38 1507.72 1646.28 1493.38 1432.62 1505.02 1554.96 1509.68
##  [41] 1512.50 1581.58 1436.42 1440.42 1387.80 1552.68 1501.22 1531.40
##  [49] 1500.24 1476.64 1566.62 1531.42 1484.20 1478.22 1466.44 1477.60
##  [57] 1587.12 1571.02 1411.46 1544.44 1500.88 1504.44 1544.20 1440.94
##  [65] 1521.24 1613.62 1473.76 1601.04 1543.08 1562.06 1368.80 1549.22
##  [73] 1601.04 1545.64 1445.58 1536.44 1386.66 1441.46 1497.68 1552.42
##  [81] 1518.58 1640.16 1513.96 1584.38 1614.00 1535.72 1540.46 1488.06
##  [89] 1345.02 1492.16 1377.50 1540.50 1524.92 1544.16 1579.82 1397.94
##  [97] 1578.54 1404.06 1419.02 1646.98
```

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

When the sample size is larger, the center converges on the true population mean, and the spread tightens.

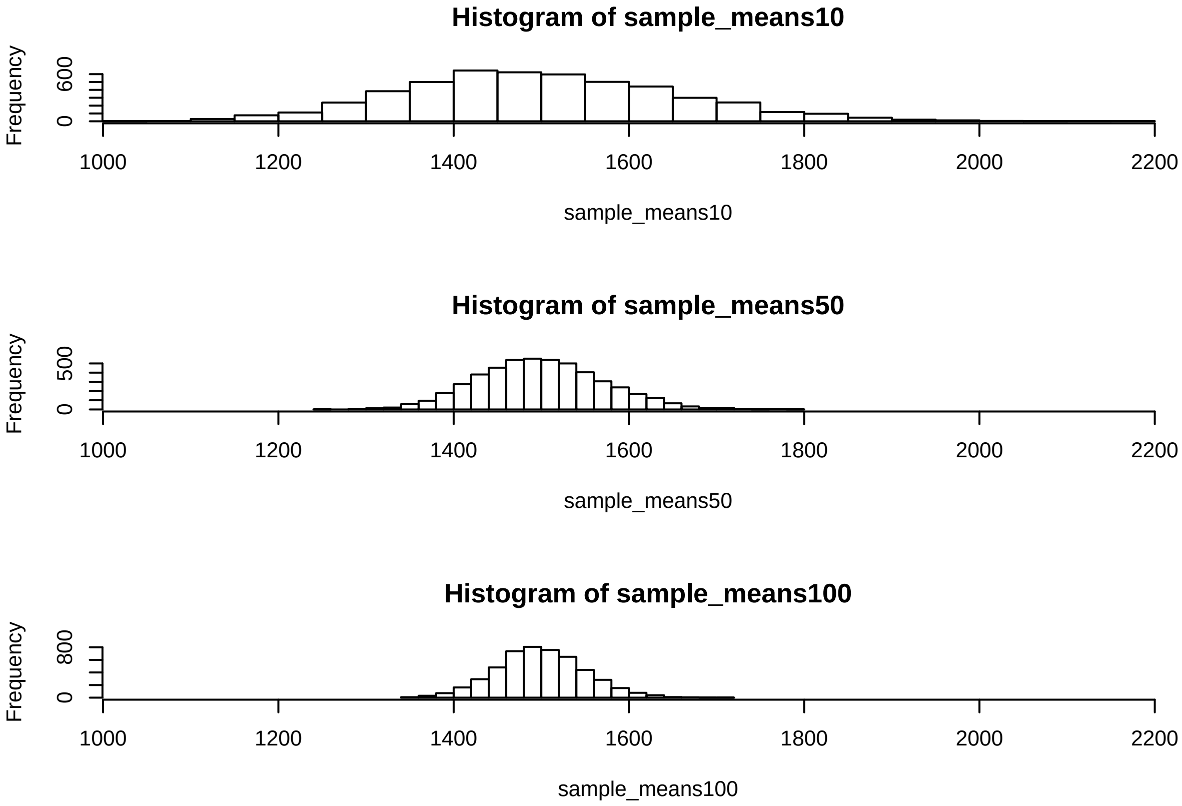
```r
sample_means10 <- rep(NA, 5000)
sample_means100 <- rep(NA, 5000)

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)
```



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?

Best point estimate of the population mean is the mean of the sample, which is 182,933.2.

```
samp3 = sample(price, 50)
mean(samp3)
```

```
## [1] 191966.6
```

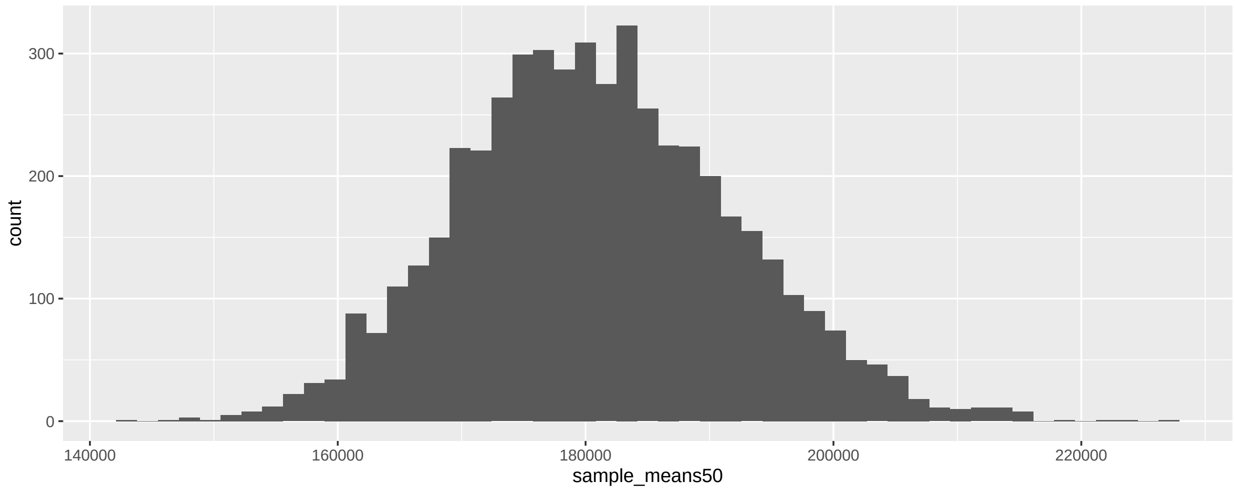
On your own: 2. Since you have access to the population, simulate the sampling distribution for x¯pricex¯price by taking 5000 samples from the population of size 50 and computing 5000 sample means. Store these means in a vector called sample_means50. Plot the data, then describe the shape of this sampling 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.

The means are normally distributed around the mean of 5000 samples of 180,973.4, which is very close to the actual population mean of 180,796.1.

```
set.seed(999)
sample_means50 <- rep(NA, 5000)

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

qplot(sample_means50, geom = 'histogram', binwidth = (max(sample_means50)-min(sample_means50))/50)
```



```
# estimated mean based on 5000 samples
mean(sample_means50)
```

```
## [1] 180973.4
```

```
# actual population mean
mean(price)
```

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

On your own: 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?

With 150 houses per sample, the distribution of means is much tighter, and the sample mean of 180,875.7 is closer to the actual population mean.

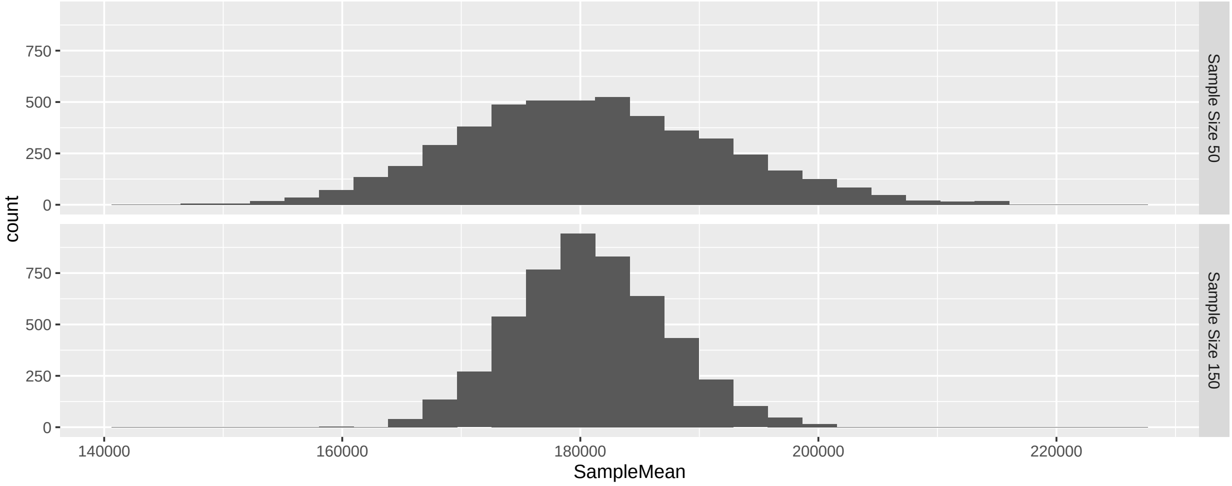
```
set.seed(999)
sample_means150 <- rep(NA, 5000)

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

# organize data to use ggplot facet_grid (while easier to use par(mfrow...), ggplot still makes prettie
# and keeps scales the same)
df50 = data.frame(rep('Sample Size 50', 5000), sample_means50)
names(df50) = c('SampleSize', 'SampleMean')
```

```
df150 = data.frame(rep('Sample Size 150', 5000), sample_means150)
names(df150) = c('SampleSize', 'SampleMean')
df = rbind(df50, df150)
df %>% ggplot(aes(x=SampleMean)) + geom_histogram() + facet_grid(SampleSize ~ .)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
# estimated mean based on 5000 samples
mean(sample_means150)
```

```
## [1] 180875.7
```

On your own: 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?

Based on the histograms above, samples of 150 houses produce a much tighter spread of sample means vs. the samples of 50 houses. As such, we would prefer a distribution with a smaller spread if we were concerned with making estimates that are more often closer to the true value.