STAT 100B Lab 2

Wesley Chang

Summer 2020 Session B

Setup for Lab

```
# set seed
require(stats)
set.seed(60)

# get data from online source
load(url("https://stat.duke.edu/~mc301/data/ames.RData"))
```

Lab Exercises

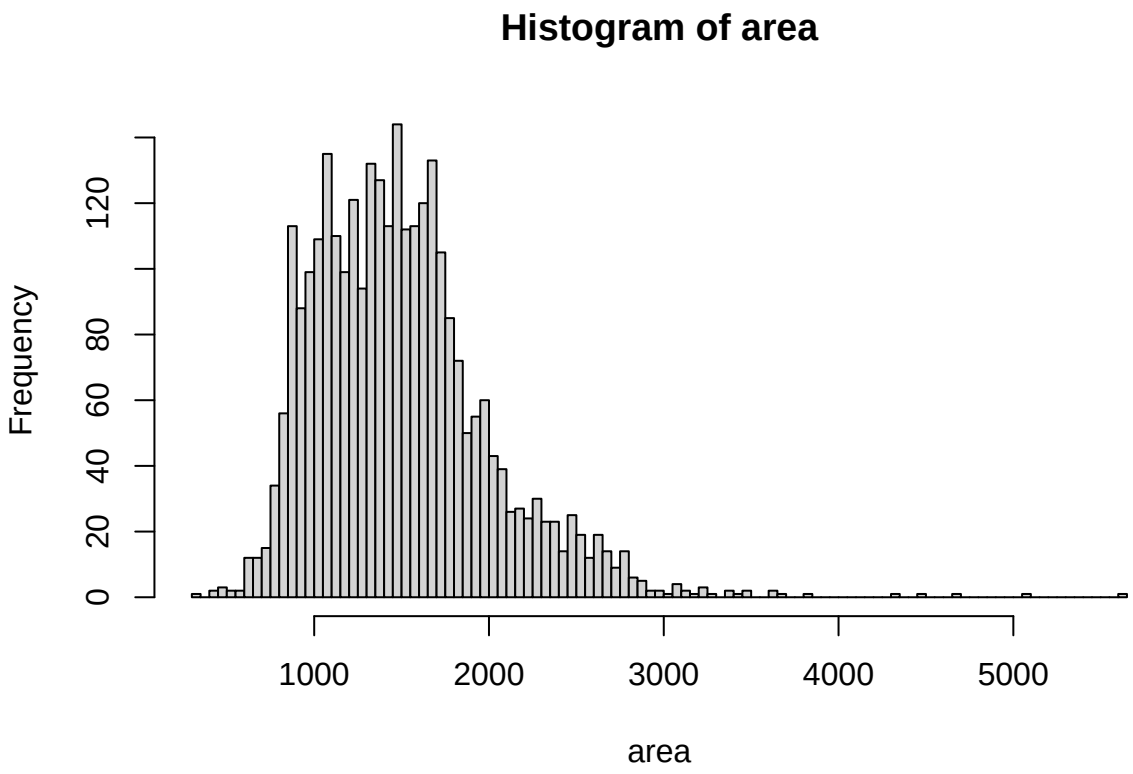
Exercise 1

Describe this population distribution using the visualization and the summary statistics. (1pt) You don't have to use all of the summary statistics in your description, you will need to decide which ones are relevant based on the shape of the distribution. Make sure to include the plot (2pts) and the summary statistics output (2pts) in your report along with your narrative.

```
# generate summary statistics for area
area <- ames$area
summary(area)
```

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

```
# generate histogram of area
hist(area, breaks = 100)
```



Answer

The population distribution has a range of 334 to 5642, with a mean of 1500 and a median of 1442. Based on the histogram, we can see that the data distribution is skewed right. This is also confirmed by the fact that the mean value is larger than the median value.

Exercise 2

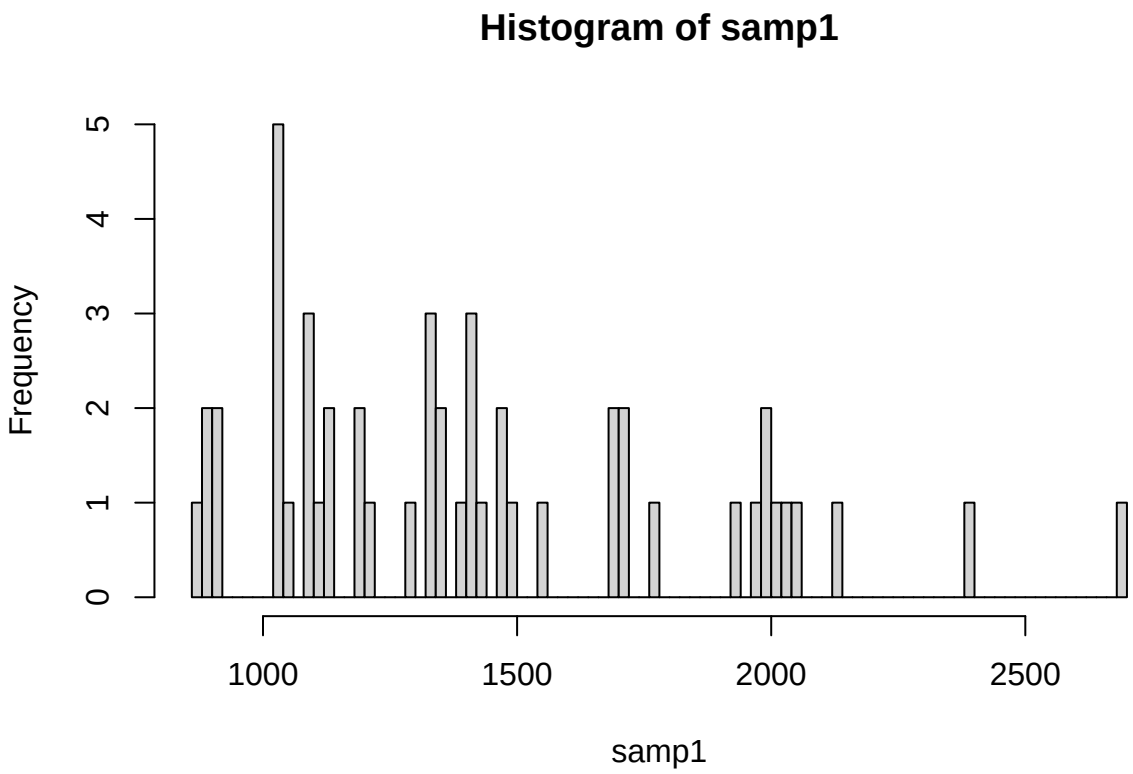
Describe the distribution of area in this sample. How does it compare to the distribution of the population? Show your code to find the mean and median of variable area in this sample.(1pt) Show your code to make a histogram of area (1pt) Distribe the distribution of area in this sample.(1pt)

```
# generate sample of population data
samp1 <- sample(area, 50)

# generate summary statistics
summary(samp1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     874    1092    1343    1436    1706    2696
```

```
# generate histogram
hist(samp1, breaks = 100)
```



Compared to the population data, we can see the the distribution shape is still skewed right, and the mean is stil larger than the median value. However, the range is much smaller, between 874 and 2696, rather than between 334 and 5642 for the population data.

Exercise 3

Would you expect the mean of your sample to match the mean of another team's sample? (1pt) Why, or why not? (1pt) If the answer is no, would you expect the means to just be somewhat different or very different?

Answer

I would not expect my mean to be *exactly* the same as mean of another team's sample. This is because sampling is likely never to exactly match the true population values. However, I would expect the mean of the other team's sample to be similar to our team's sample mean.

Exercise 4

Take a second sample, also of size 50, and call it samp2. How does the mean of samp2 compare with the mean of samp1? (1pt) Show your code to build samp2 and find the mean of samp2. (2pts)

```
# generate samp2
samp2 <- sample(area, 50)
# summary statistics of samp2
summary(samp2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     492    1191    1412    1457    1708    2541
```

Answer

The mean values are really similar, with 1457 for samp2 vs 1436 for samp2.

Exercise 5

Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimation of the population mean?

Answer

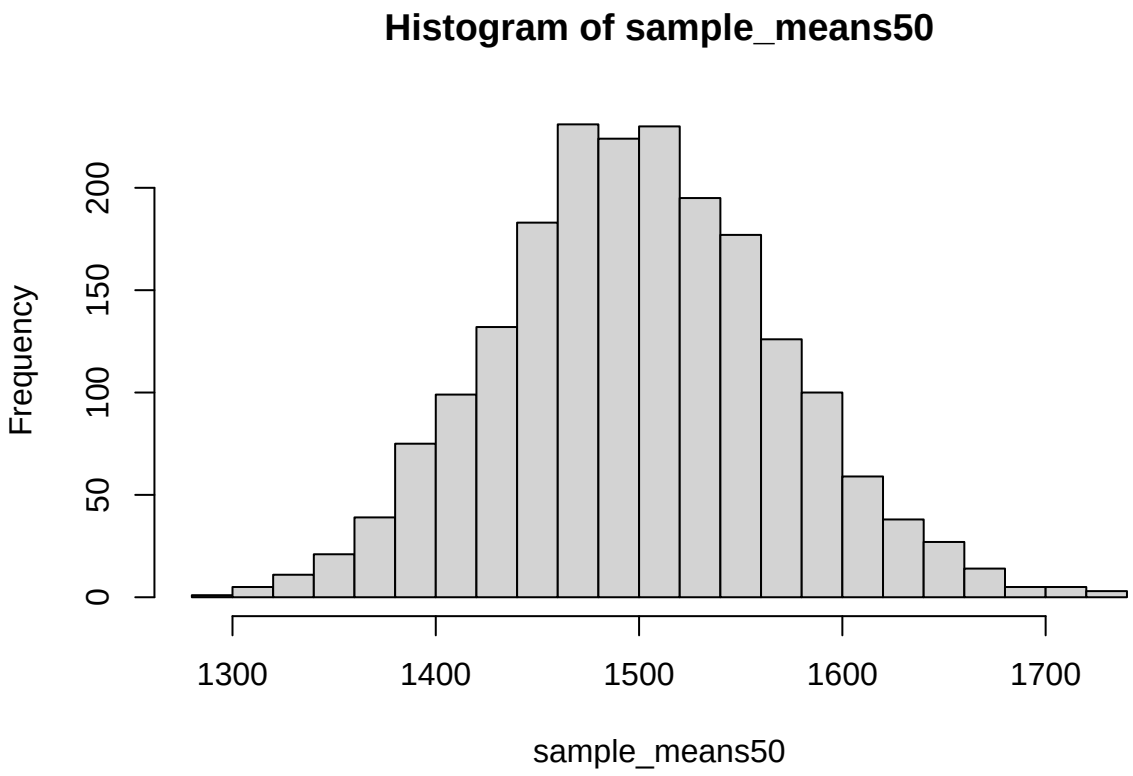
By the Central Limit Theorem, I would assume that the sample with n = 1000 would provide a more accurate estimation of the population mean

Exercise 6

How many elements are there in sample_means50? (1pt) Describe the shape of the sampling distribution (1pt) The sampling distribution is centered at_____?

```r
# build an empty vector to hold sample means
num_samples <- 2000
sample_means50 <- rep(0, num_samples)

# generate 2000 samples of size 50
# calculate sample means and store them in vector sample_mean50
for (i in 1:num_samples){
  temp_samp <- sample(area, 50)
  sample_means50[i] <- mean(temp_samp)
}
# visualize the sampling distribution
hist(sample_means50, breaks = 20)
```



There are 2000 different sample means, each of sample size 50 in the sample_means50. The sample distribution is fairly symmetrical, with a a center at about 1500.

On your own (Lab B Exercises)

Question 1

Take a sample of size 15 from the population and calculate the mean price of the homes in this sample. Show your code to obtain the mean of this sample (1pt) Using this sample, what is your best point estimate of the population mean of prices of homes? (1pt)

```
price <- ames$price
# Take a sample of size 15
q1 <- sample(ames$price, 15)

# find mean price of homes in this sample
q1mean <- mean(q1)
q1mean
```

```
## [1] 207421.8
```

Answer

Using this sample, my best point estimate of the population mean of the prices of homes is 191753.3.

Question 2

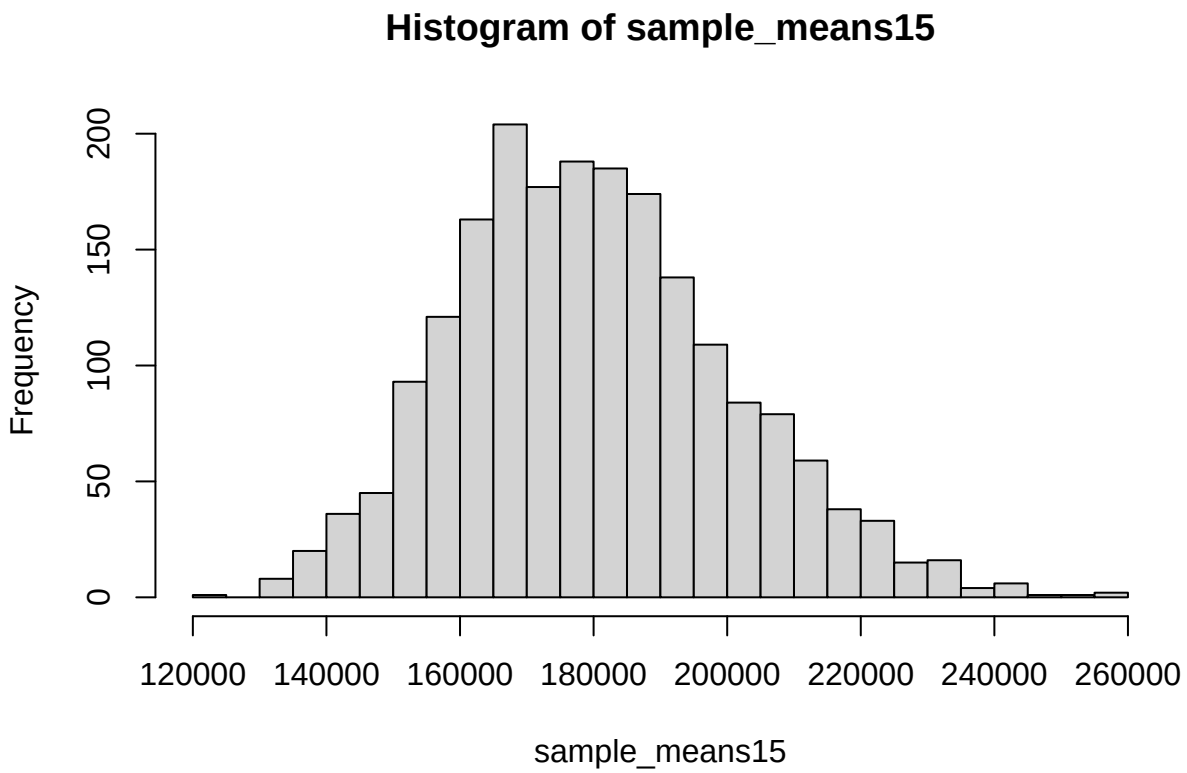
Since you have access to the population, simulate the sampling distribution for $\bar{x}$price by taking 2000 samples from the population of size 15 and computing 2000 sample means. Store these means in a vector called sample_means15. Show your code of the sampling process.(2pts) Show your code of plotting sampling distribution of size 15. (1pt) Describe the shape of this sampling distribution (1pt) Finally, show your code to calculate and report the population mean.(2pts)

```
# simulate the sampling distribution for xhat price
# take 2000 samples from the pop of size 15
# num_samples was defined earlier as 2000

# create empty vector to hold sample means
sample_means15 <- rep(0, 2000)

# generate 2000 samples of size 15
# calculate sample means and store them in vector sample_mean15
for (i in 1:2000){
    temp_samp15 <- sample(price, 15)
    sample_means15[i] <- mean(temp_samp15)
}

# Plot sampling distribution in histogram
hist(sample_means15, breaks = 20)
```



```r
# code to calculate the population mean
mean(price)
```

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

```r
# population mean from 2000 sampling distributions
mean(sample_means15)
```

```
## [1] 180011.8
```

Answer

The distribution of the sampling means of size 15 looks relatively symmetrical when plotted on a histogram. The population mean is 180796.1, while the mean of the 2000 sample means is 180365.4.

Question 3

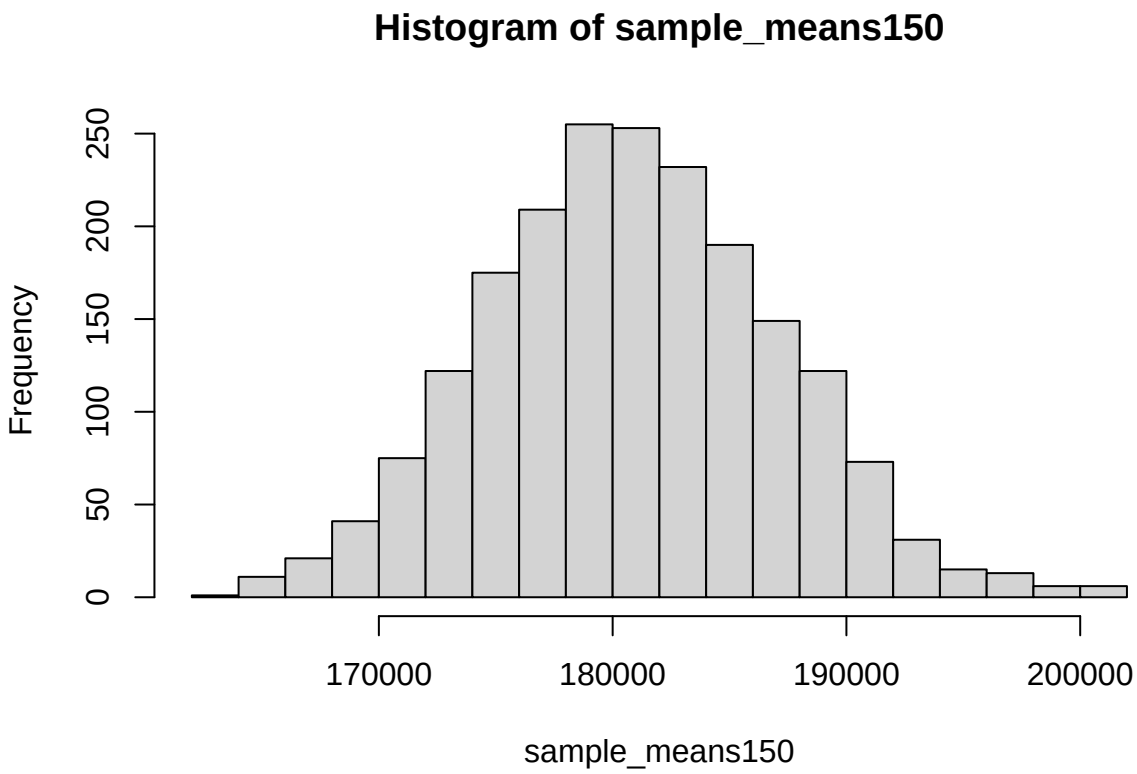
Change your sample size from 15 to 150, then compute the sampling distribution using the same method as above, and store these means in a new vector called sample_means150. Show your code here for sampling and ploting (4pts) Comparing to the sampling distribution from a sample size of 15, the spread of this sampling distribution is _____ (smaller/larger) when sample size increased to 150.(1pt) The shape of this sampling distribution is _____(1pt) A. roughly bell shaped B.not close to bell shaped

```r
# distribution of 2000 sample means of sample size 150

# create empty vector
sample_means150 <- rep(0,2000)

# generate 2000 samples of size 150
for (i in 1:2000){
    temp_samp150 <- sample(price, 150)
    sample_means150[i] <-mean(temp_samp150)
}

# plot histogram of sample_means150
hist(sample_means150, breaks = 20)
```



Answer

Compared the sampling distribution of sample size 15, the spread of this sampling distribution is smaller when the sample size increased to 150. The shape of this sampling distribution is roughly bell-shaped.

Question 4

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? (1pt)

Answer

If we're concerned with making estimates that are more often close to the true value, we would prefer a sampling distribution with a smaller spread.