lab4

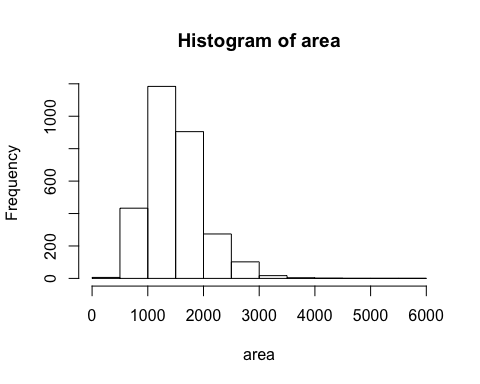
jason grahn

10/15/2018

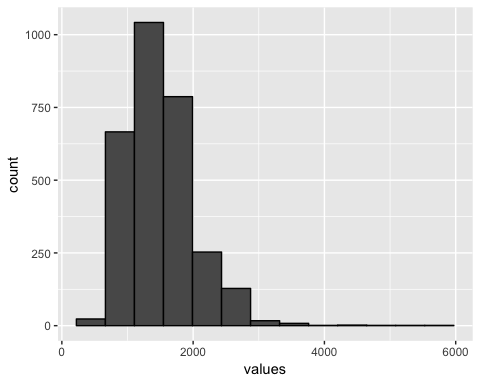
#base R  
area <- ames$Gr.Liv.Area  
price <- ames$SalePrice  
   
summary(area)

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

hist(area)



#tidy-style  
ames %>%   
 select(area = Gr.Liv.Area,   
 price = SalePrice) %>%   
 gather(key = 'variable', value = 'values', area, price) %>%   
 filter(variable == "area") %>%   
 ggplot() +  
 geom\_histogram(aes(x = values),   
 bins = 13,  
 color = "black")



# Exercise1

## Describe this population distribution.

The distribution appears unimodal with a long right skew. The median and mean quite close at 1442 and 1500, respectively, which indicates some degree of normalicy.

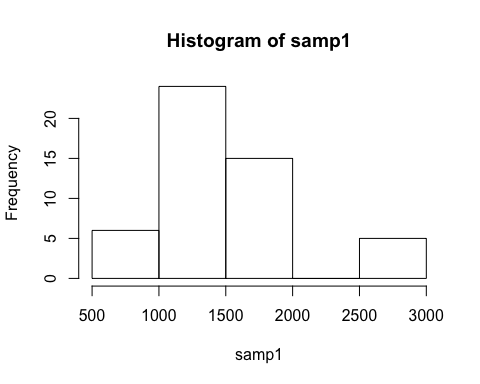
# Excercise2

## Describe the distribution of this sample. How does it compare to the distribution of the population?

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

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 747 1114 1354 1467 1582 2810

hist(samp1)



The Mean and Median of the sameple is similar, albeit different than the population. The histogram shows similar tendencies, unimodal with right skew.

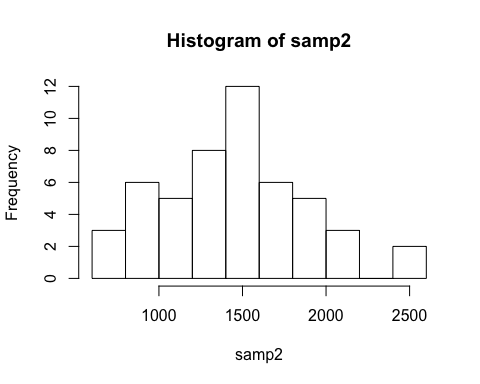
# Exercise3

## 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?

samp2 <- sample(area, 50)  
  
summary(samp2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 630 1148 1489 1455 1683 2589

hist(samp2)



The mean of this sample is still similar to sample 1, though a bit higher. If we took larger samples of 100 and 1000, I would expect the sample means to be much closer to the population mean.

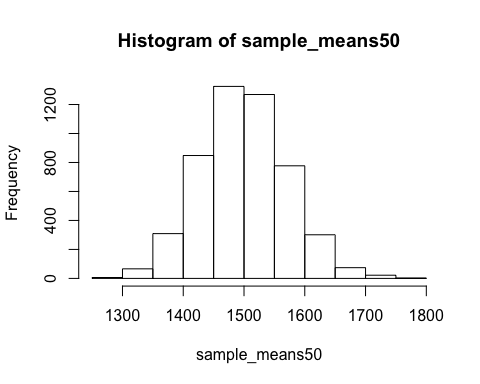
# Exercise4

## 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?

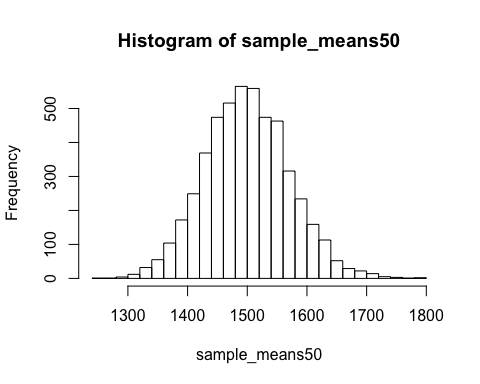
sample\_means50 <- rep(NA, 5000)  
  
for(i in 1:5000){  
 samp <- sample(area, 50)  
 sample\_means50[i] <- mean(samp)  
 }  
  
summary(sample\_means50)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1257 1451 1498 1500 1547 1784

hist(sample\_means50)



#If you would like to adjust the bin width of your histogram to show a little more detail, you can do so by changing the breaks argument.  
  
hist(sample\_means50, breaks = 25)



Sample\_means50 has 5000 elements. The sampling distribution appears quite normal with a mean of 1499 and median 1497. Given we’re so incredibly far past the point of central limit theory, I would *not* expect the distribution to change if we collected 50,000 instead. It might be a little taller in the center, but otherwise unchanged.

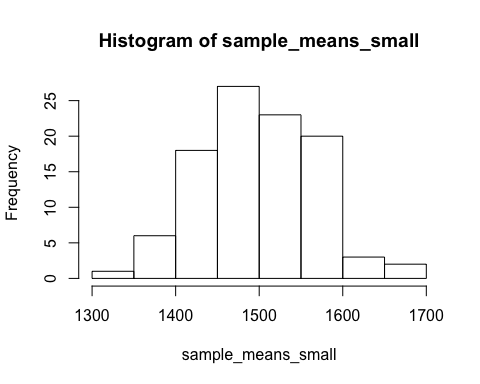
# Example5

## 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?

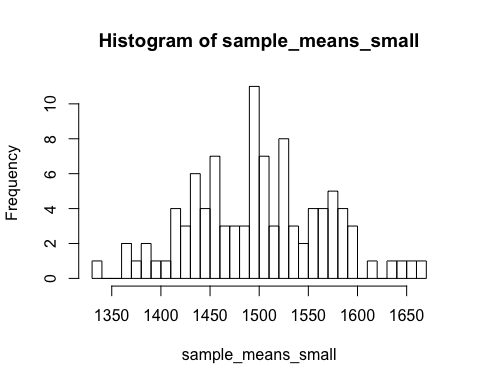
sample\_means\_small <- rep(NA, 100)  
  
for(i in 1:100){  
 samp <- sample(area, 50)  
 sample\_means\_small[i] <- mean(samp)  
 }  
  
summary(sample\_means\_small)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1337 1451 1498 1499 1550 1669

hist(sample\_means\_small)



#If you would like to adjust the bin width of your histogram to show a little more detail, you can do so by changing the breaks argument.  
  
hist(sample\_means\_small, breaks = 25)



print(sample\_means\_small)

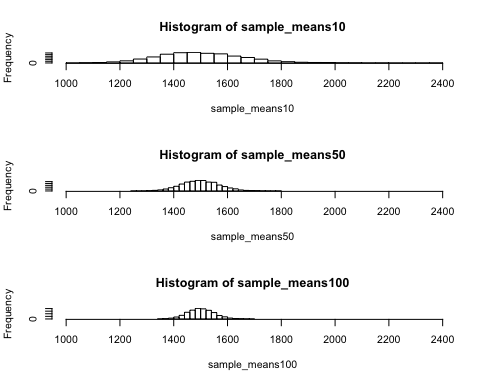
## [1] 1523.48 1528.16 1407.46 1561.26 1586.04 1587.84 1564.00 1458.02  
## [9] 1519.54 1396.58 1443.30 1549.48 1492.06 1501.56 1447.00 1454.74  
## [17] 1415.12 1459.24 1459.64 1436.56 1429.06 1498.22 1579.94 1375.06  
## [25] 1497.12 1410.82 1500.24 1617.54 1496.88 1502.66 1457.98 1596.34  
## [33] 1557.90 1492.10 1528.44 1658.46 1488.66 1431.22 1511.88 1573.56  
## [41] 1481.56 1642.06 1492.10 1426.18 1578.20 1477.16 1492.66 1536.42  
## [49] 1504.24 1520.78 1502.80 1449.68 1472.44 1467.38 1434.56 1567.80  
## [57] 1509.12 1524.54 1430.68 1499.02 1455.64 1478.38 1433.06 1451.34  
## [65] 1389.82 1443.62 1591.62 1560.30 1631.86 1669.28 1538.82 1537.58  
## [73] 1529.88 1514.00 1434.64 1484.90 1551.66 1429.14 1500.86 1368.58  
## [81] 1570.32 1368.08 1551.62 1336.96 1580.44 1461.72 1491.86 1413.56  
## [89] 1469.54 1387.72 1544.74 1492.72 1596.44 1524.62 1491.54 1523.44  
## [97] 1587.66 1553.90 1579.04 1414.08

There are 100 elements in the vector sample\_means\_small and each element represents the mean of the size-50 sample that was taken in the for loop.

# Exercise6

## When the sample size is larger, what happens to the center? What about the spread?

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)



When the sample size is larger, the center gets taller and the spread gets smaller.