Exercise4 - Foundations for statistical inference

Shikhar Dhwaj

29 September 2017

Reading the data file

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

Select the relevent variables

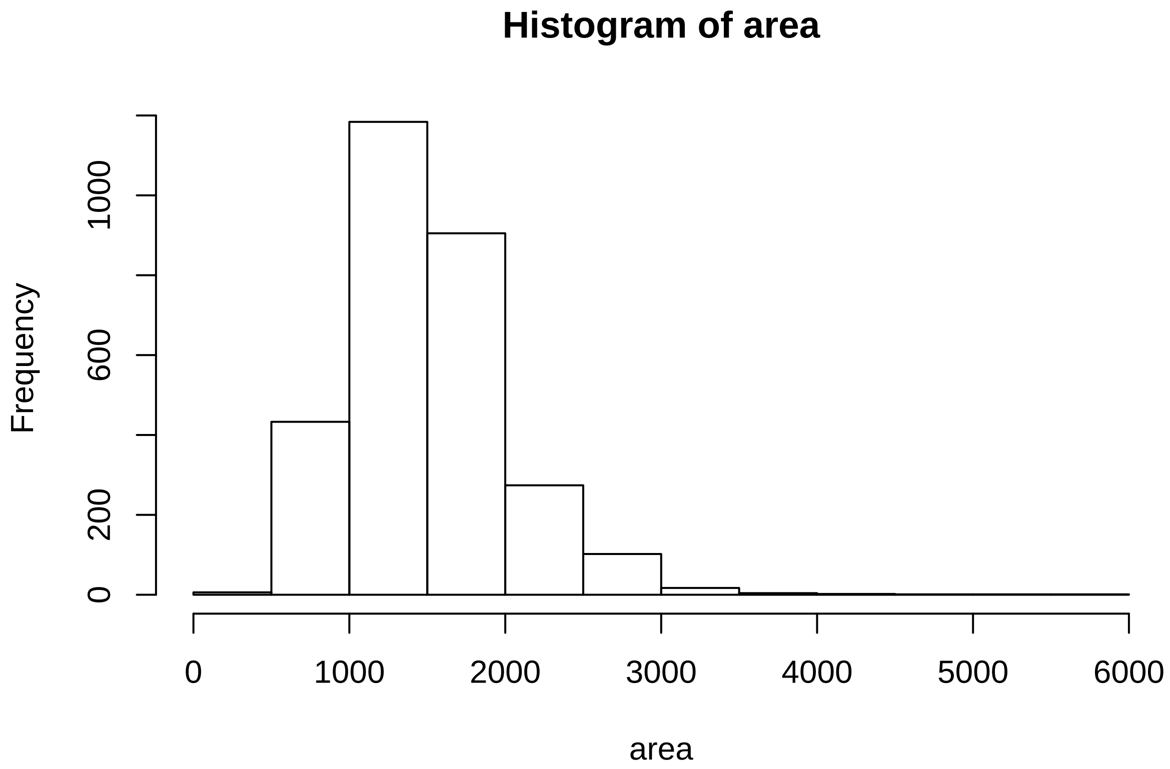
```
area <- ames$Gr.Liv.Area
price <- ames$SalePrice
```

Q-1)Describe this population distribution.

```
summary(area)
```

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

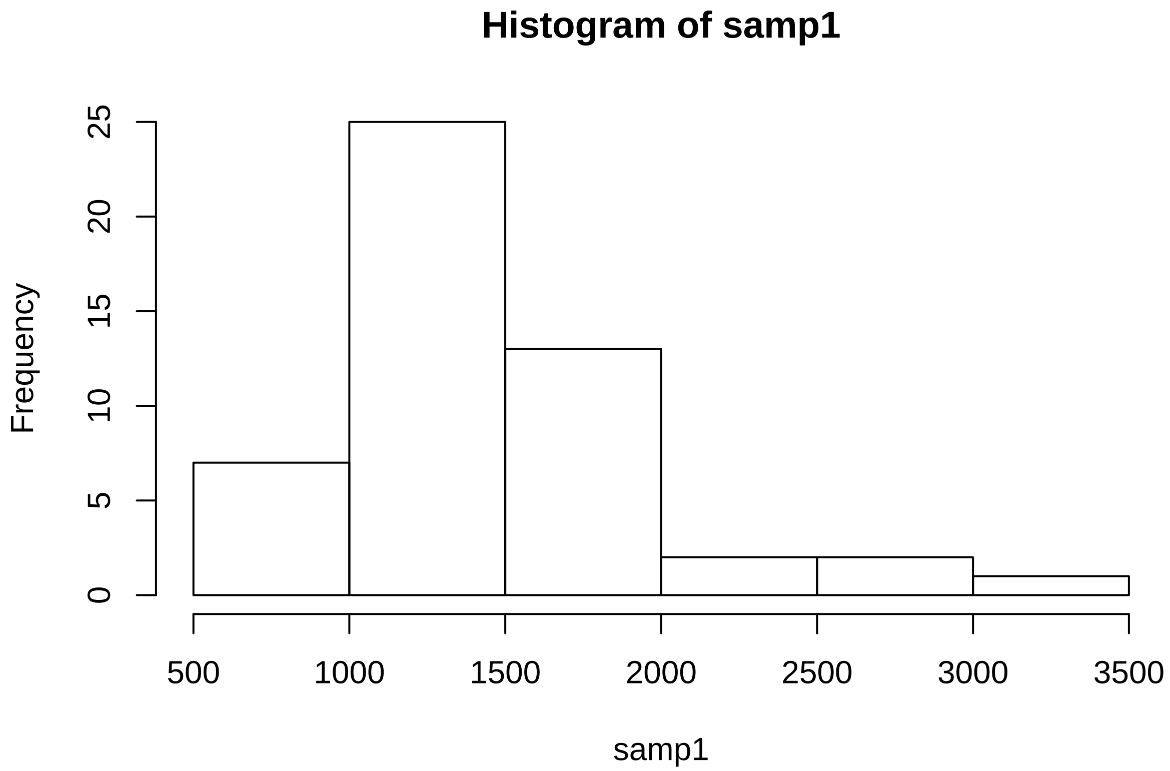
```
hist(area)
```



The distribution of areas of houses in Ames is unimodal and right-skewed. The middle 50% of the houses range between approximately 1,130 square feet and 1,740 square feet. The IQR is approximately 610 square feet. The smallest house is 334 square feet and the largest is 5,642 square feet.

Q-2)Describe the distribution of this sample. How does it compare to the distribution of the population?

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



We're interested in estimating the average living area of homes in Ames using the sample, our best single guess is the sample mean

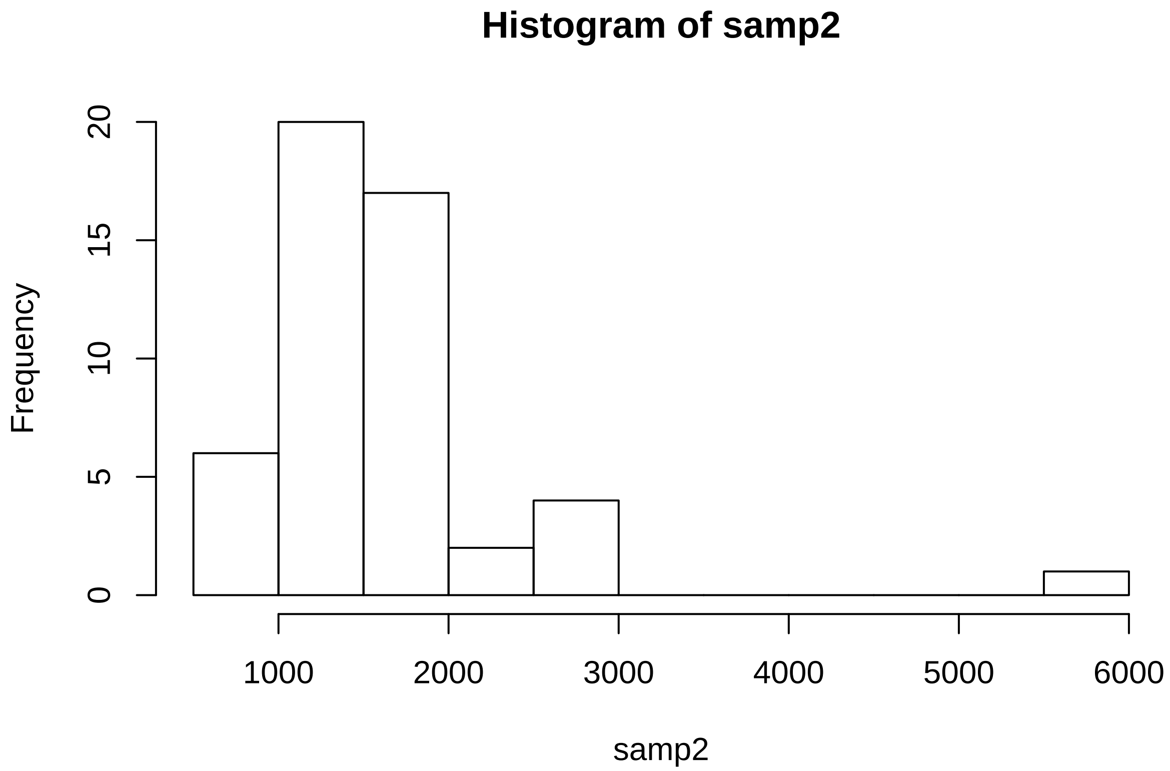
```
mean(samp1)
```

```
## [1] 1455.6
```

Population Mean is 1500

Q-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?

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



```
cat("Mean of sample2 = ",mean(samp2))
```

```
## Mean of sample2 =  1618.94
```

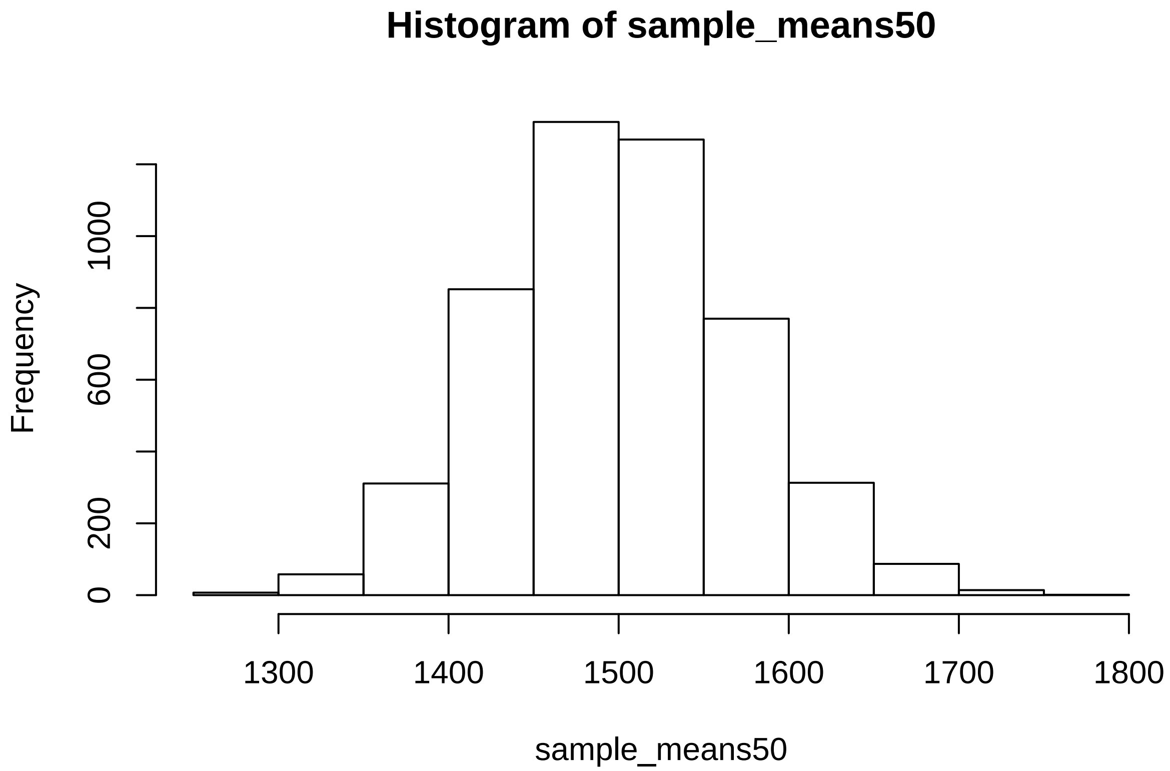
```
cat("Mean of sample1 = ",mean(samp1))
```

```
## Mean of sample1 =  1455.6
```

Size of 1000 would provide more accurate estimate of population mean, as sample size is inversly proportional to sampling error

Q-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?

```
sample_means50 <- rep(NA, 5000)
for (i in 1:5000) {
  samp <- sample(area, 50)
  sample_means50[i] <- mean(samp)
}
hist(sample_means50, breaks = 13)
```



```
cat("Mean = ",mean(sample_means50))
```

```
## Mean =  1500.202
```

There are 5,000 elements in sample_means50. The sampling distribution is normal and unimodal with a center of approximately 1500. Even though additional samples help identify the actual population mean, the sampling distribution will not change from unimodal normal

Q-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. 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)
}
length(sample_means_small)
```

```
## [1] 100
```

There are 100 elememts in sample_means_small Each element represents a mean square footage from a simple random sample of 50 houses.

Q-6)When the sample size is larger, what happens to the center? What about the spread?

As 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 decreses.

Q-7)Take a random sample of size 50 from price. Using this sample, what is your best point estimate of the population mean?

```
sample_50 <- sample(price,size=50)
mean(sample_50)
```

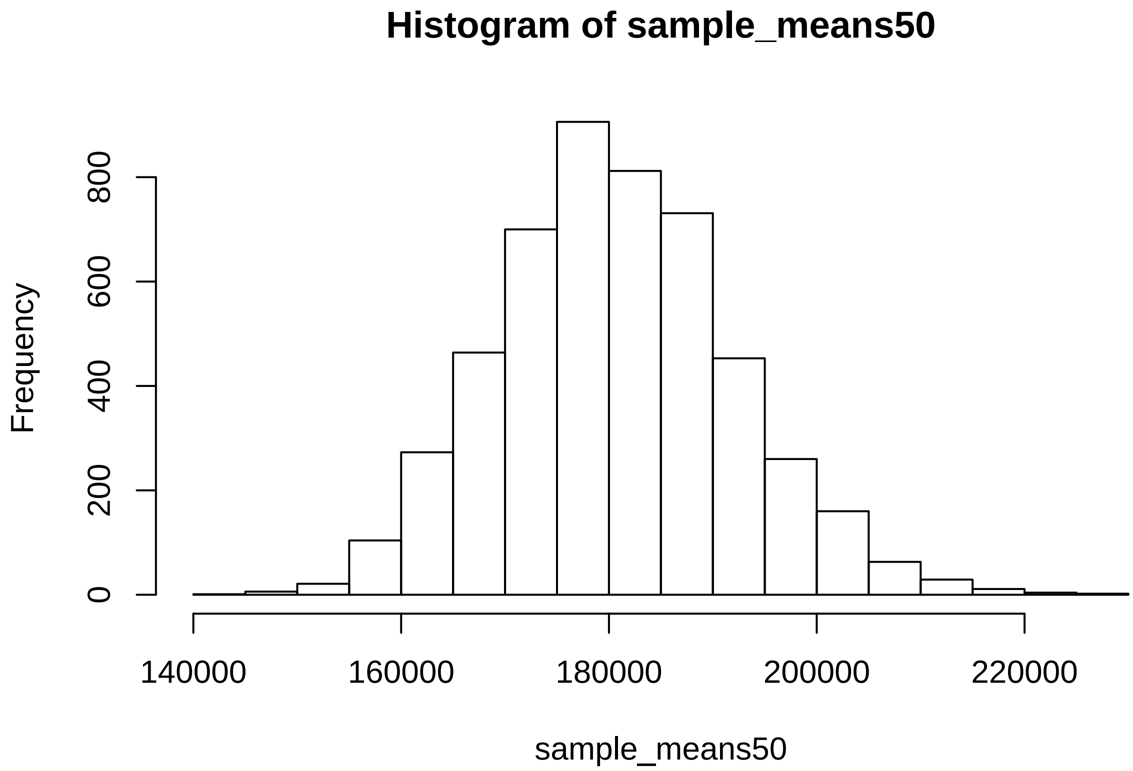
```
## [1] 175820.8
```

Q-8)Since you have access to the population, simulate the sampling distribution for 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,size=50)
sample_means50[i] <- mean(samp)
}
summary(sample_means50)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  144087  172970  180137  180679  187995  226406
```

```
hist(sample_means50)
```



```
summary(price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12789  129500  160000  180796  213500  755000
```

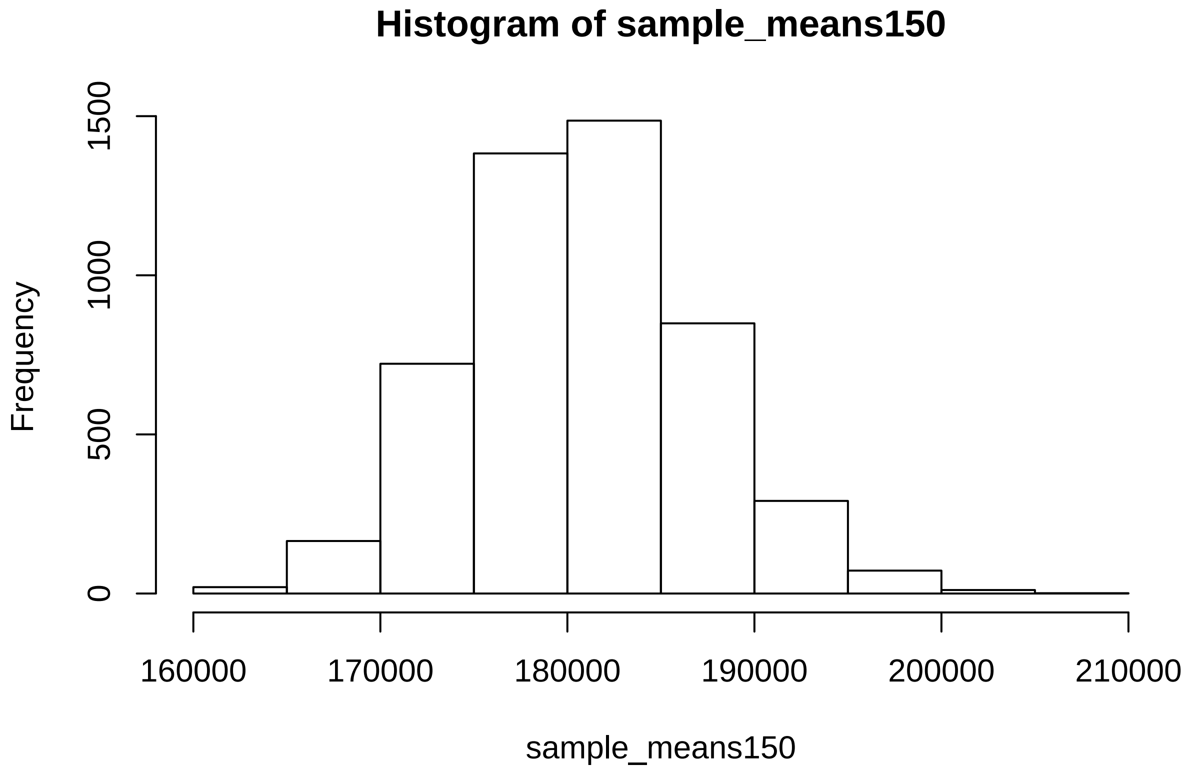
The shape of the sampling distribution is normal and unimodal. Using the sampling distribution, I would guess that the population's mean would be 180,796.

Q-9)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,size=150)
sample_means150[i] <- mean(samp)
}
summary(sample_means150)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  160237  176387  180728  180807  184896  205853
```

```r
hist(sample_means150)
```



```r
cat("Sale price estimated is ",mean(sample_means150))
```

```
## Sale price estimated is  180806.6
```

The variability of the sampling distribution with the smaller sample size (sample_means50) is larger than the variability of the sampling distribution with the larger sample size (sample_means150). The means for the two sampling distributions are roughly similar.

Q-10)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?

Of the two sampling distributions (sample_means50 and sample_means150), the second (sample_means150) has the smaller spread. It is preferable to have a distribution with a smaller spread when trying to make estimates that are close to the true value, because there is less uncertainty of where the true estimates lie.