Sampling Distributions

Debopriya Ghosh

2019-09-13

In this lecture, we will investigate the ways in which the statistics from a random sample of data can serve as point estimates for population parameters. We're interested in formulating a sampling distribution of our estimate in order to learn about the properties of the estimate, such as its distribution.

Importing the data

We will use the real estate data from the city of Ames, Iowa. The details of every real estate transaction in Ames is recorded by the City Assessor's office. We will focus only on all residential home sales in Ames between 2006 and 2010. This collection represents our population of interest. In this exercise we would like to learn about these home sales by taking smaller samples from the full population.

```
ames = read.csv("~/Dropbox/Priya-PhD- Documents/Courses/Data Analysis and Visualization-Fall 2019/Datas
```

We see that there are 81 variables in the data set. For our analysis, we'll restrict our attention to just two of the variables: the above ground living area of the house in square feet (Gr.Liv.Area) and the sale price (SalePrice).

```
area = ames$GrLivArea
price = ames$SalePrice
```

Let's look at the distribution of area in our population of home sales by calculating a few summary statistics and making a histogram.

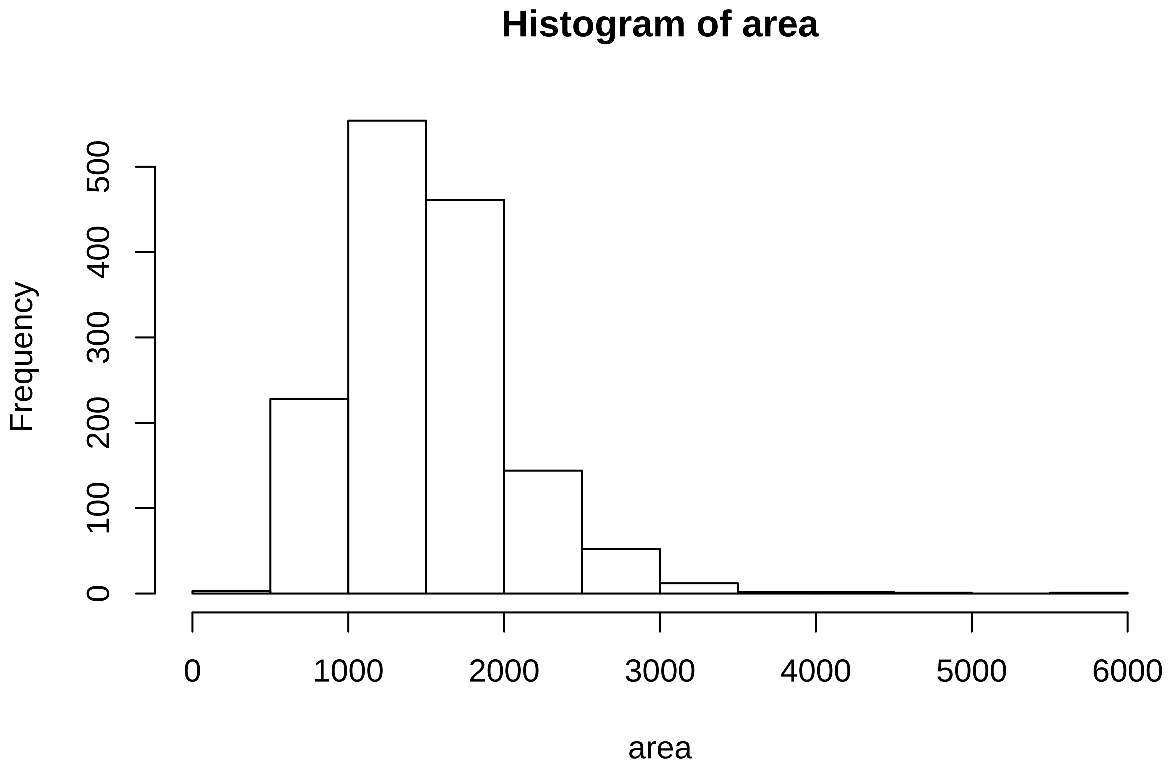
```
summary(area)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     334    1130    1464    1515    1777    5642
```

```
mean(area)
```

```
## [1] 1515.464
```

```
hist(area)
```



Describing the population

The distribution of area is right skewed. It is unimodel. Mean is around 1515. And SD is of 525.4804

The unknown sampling distribution

Assume here we have access to the entire population, but this is rarely the case in real life. Gathering information on an entire population is often extremely costly or impossible. Because of this, we often take a sample of the population and use that to understand the properties of the population.

If we were interested in estimating the mean living area in Ames based on a sample, we can use the following command to survey the population.

```
samp1 = sample(area, 50)
```

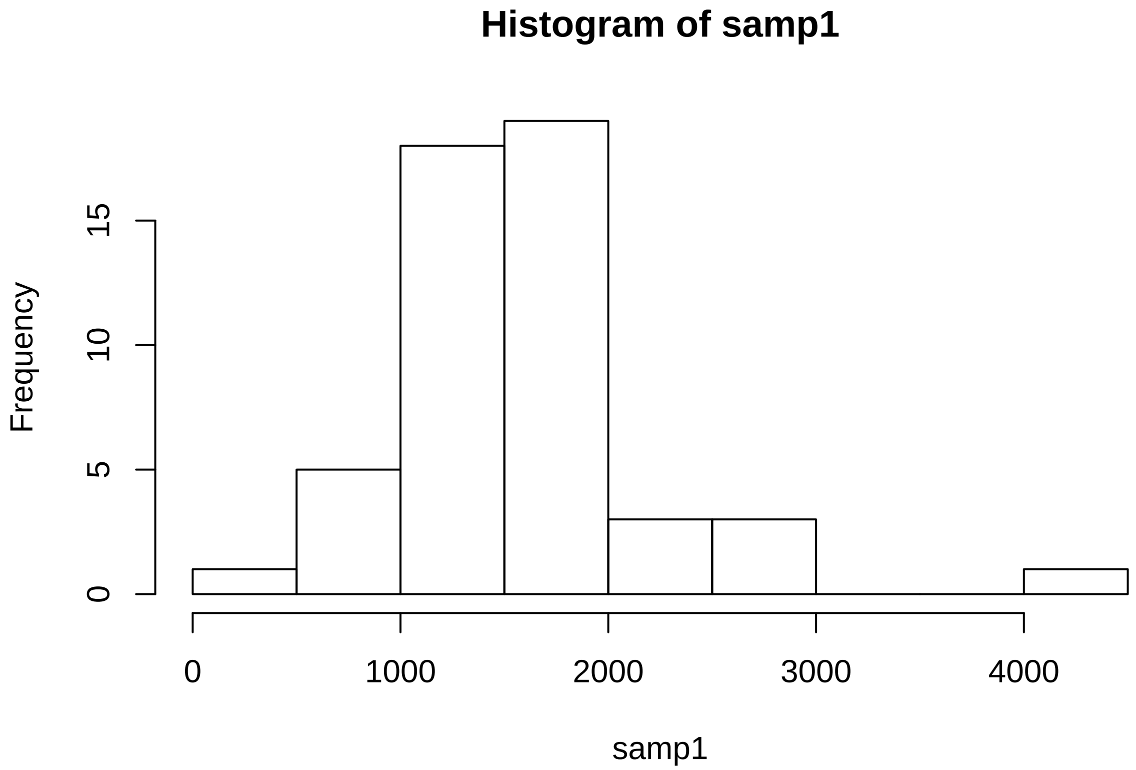
This collects a simple random sample of size 50 (50 random home sales) from the vector area.

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

```
summary(samp1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     334    1181    1538    1556    1732    4316
```

```
hist(samp1)
```



This sample distribution is little bit right skewed(It may vary on each sample). It is has a mean of 1561.22. As the sample size is large enough, we can assume central limit therom for this sample.

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

```
mean(samp1)
```

```
## [1] 1556.1
```

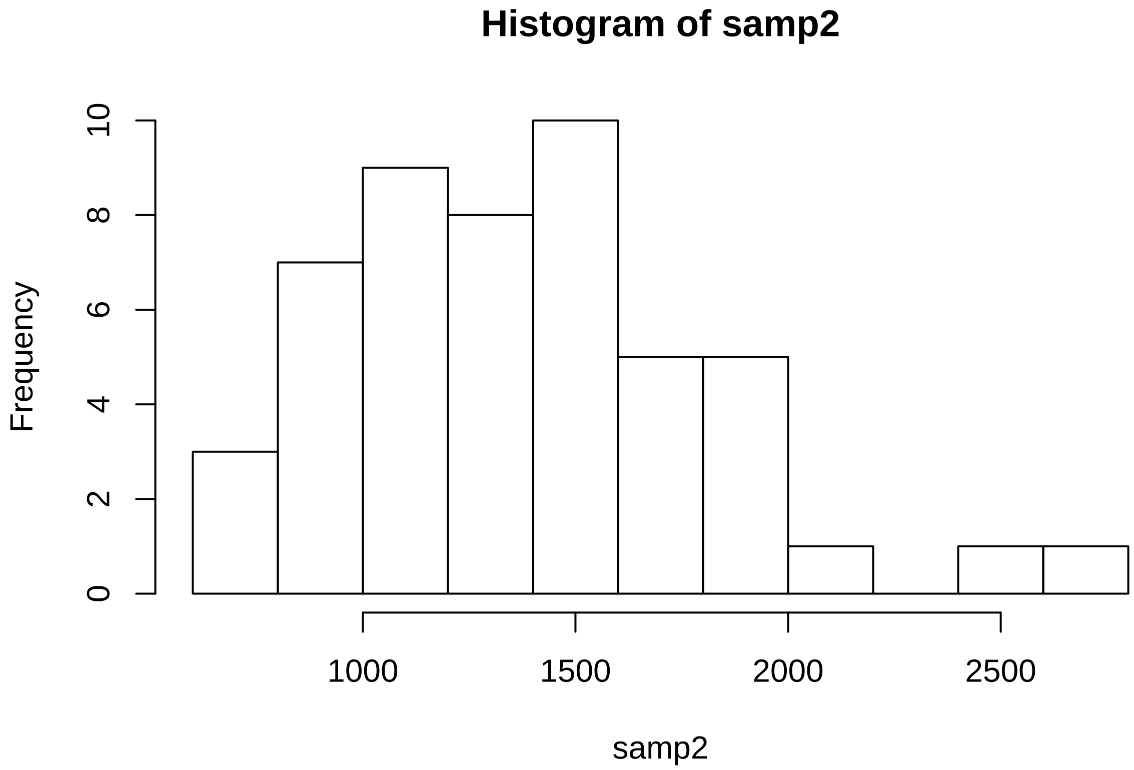
Depending on which 50 homes you selected, your estimate could be a bit above or a bit below the true population mean of 1515.464 square feet. In general, though, the sample mean turns out to be a pretty good estimate of the average living area, and we were able to get it by sampling around 3% of the population.

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)
mean(samp2)
```

```
## [1] 1396.44
```

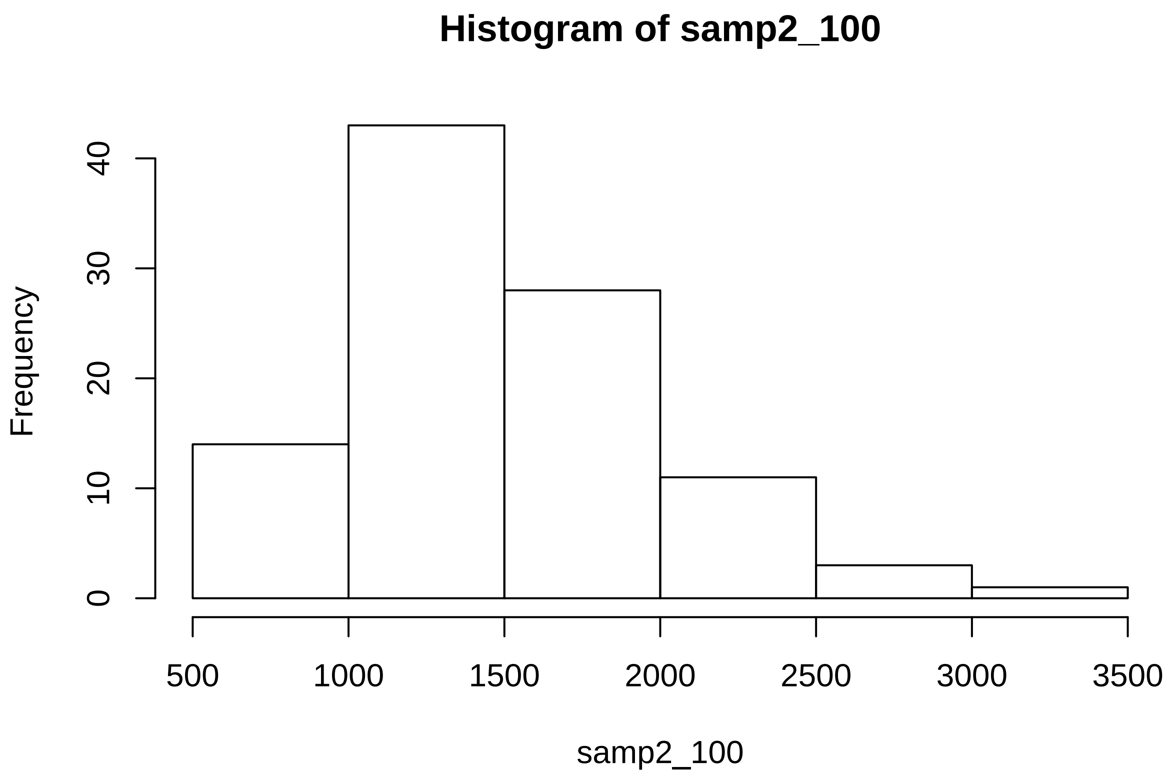
```r
hist(samp2)
```



```r
samp2_100 = sample(area,100)

mean(samp2_100)
```

```
## [1] 1504.47
```

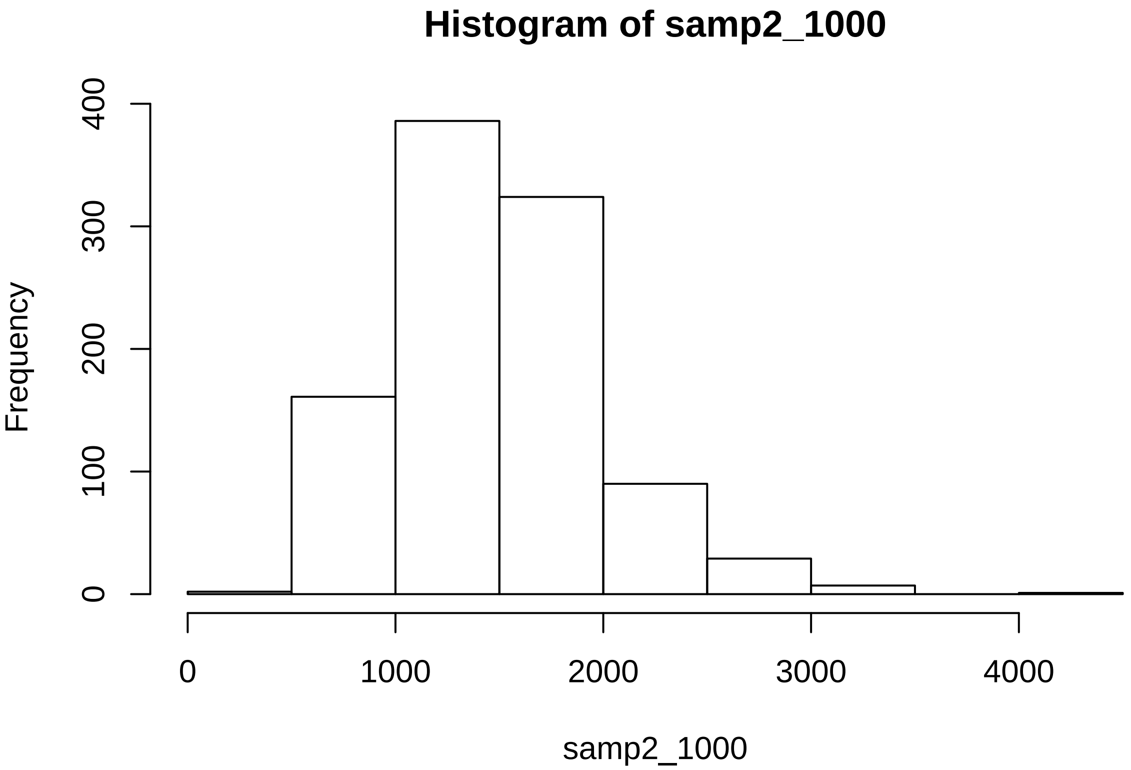
```r
hist(samp2_100)
```



```
samp2_1000 = sample(area,1000)

mean(samp2_1000)
```

```
## [1] 1487.953
```

```
hist(samp2_1000)
```

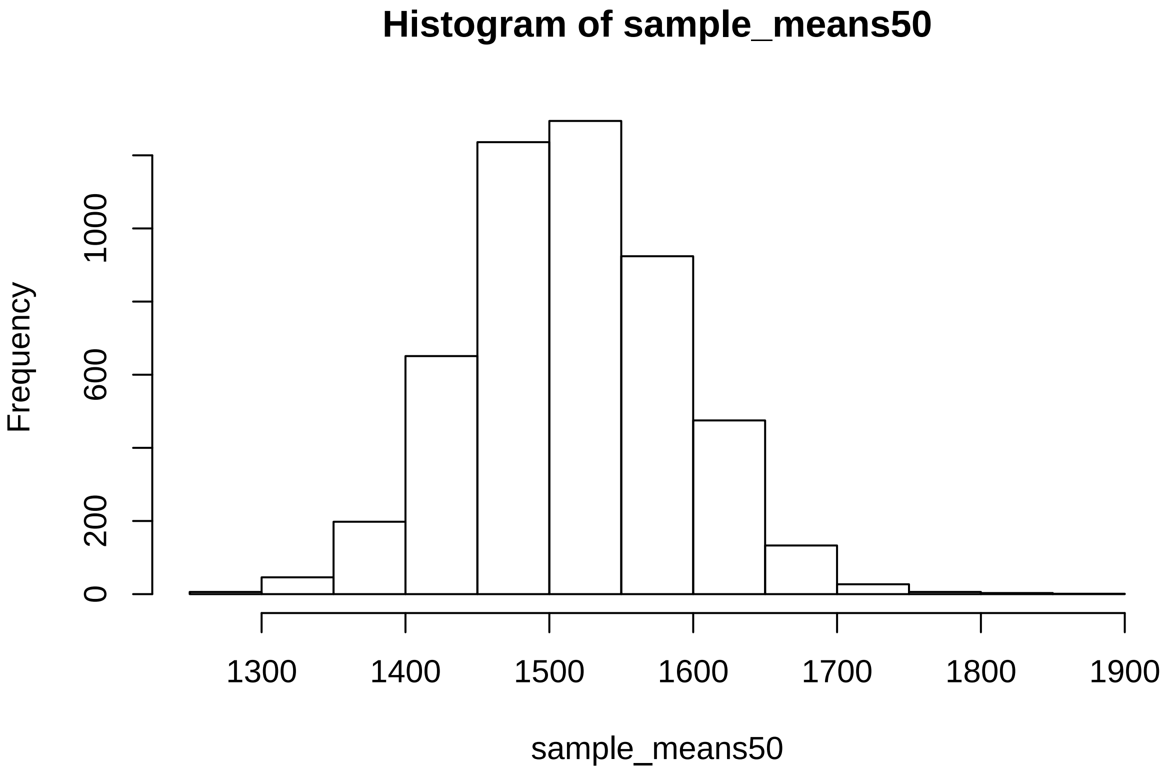


Above samples show that more the number of sample is good for accurate estimate of the population. But to ensure that sample observations are independet, we need to conduct simple random sample consisting of less than 10% of the population. So in that scenario, 100 might be a good sample size.

Every time we take a random sample, we get a different sample mean. Therefore, it's useful to get a sense of just how much variability we should expect when estimating the population mean this way. The distribution of sample means, called the sampling distribution, can help us understand this variability.

Here we have access to the population, we can build up the sampling distribution for the sample mean by repeating the above steps many times. We will generate 5000 samples and compute the sample mean of each.

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

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

hist(sample_means50)
```

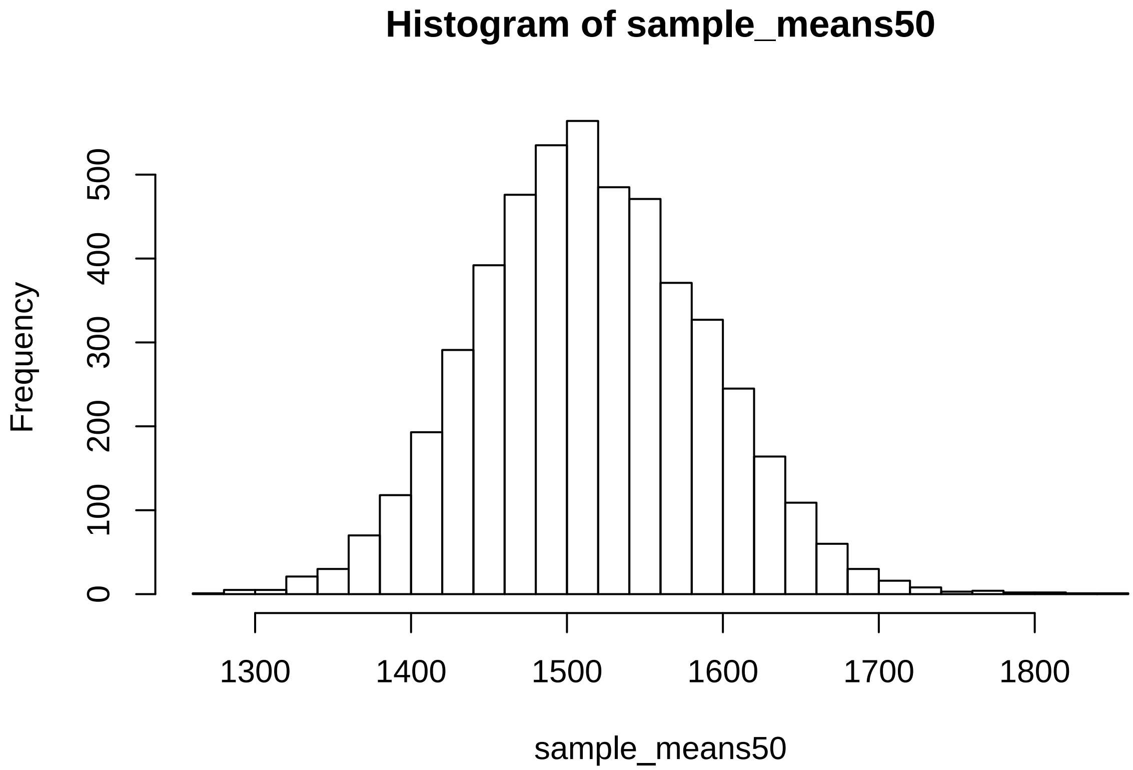


```r
mean(sample_means50)
```

```
## [1] 1515.864
```

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.

```r
hist(sample_means50, breaks = 25)
```



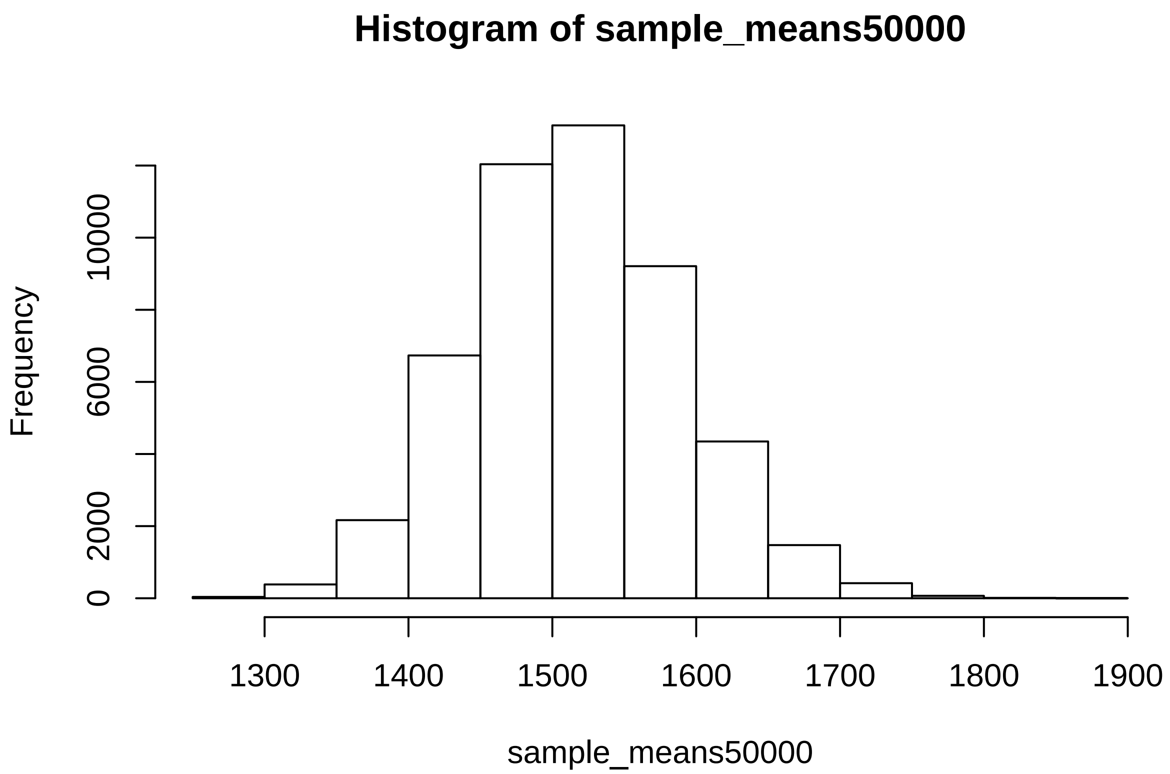
Here we take 5000 samples of size 50 from the population, calculate the mean of each sample, and store each result in a vector called *sample_means50*.

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 contains 5000 elements. The mean is 1501.322152. It is almost similar to the original dataset.

The sample mean will not(almost) will not change if we have 50,000 sample means.

```
sample_means50000 = rep(NA, 50000)

for(i in 1:50000){
   samp = sample(area, 50)
   sample_means50000[i] = mean(samp)
   }

hist(sample_means50000)
```



```
mean(sample_means50000)
```

```
## [1] 1515.709
```

Sample size and the sampling distribution

The sampling distribution tells us much about estimating the average living area in homes in Ames. Because the sample mean is an unbiased estimator, the sampling distribution is centered at the true average living area of the the population, and the spread of the distribution indicates how much variability is induced by sampling only 50 home sales.

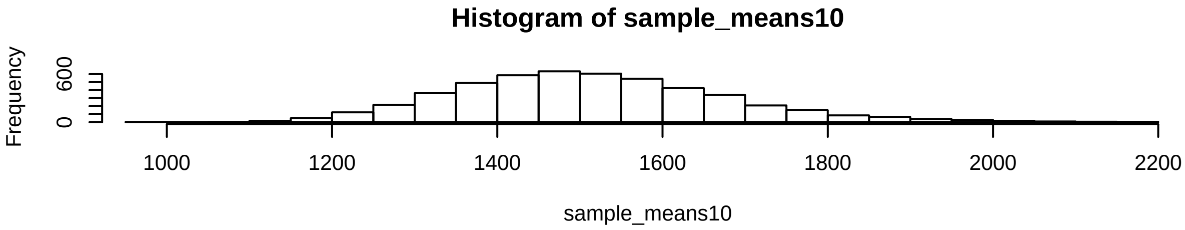
To get a sense of the effect that sample size has on our distribution, let's build up two more sampling distributions: one based on a sample size of 10 and another based on a sample size of 100.

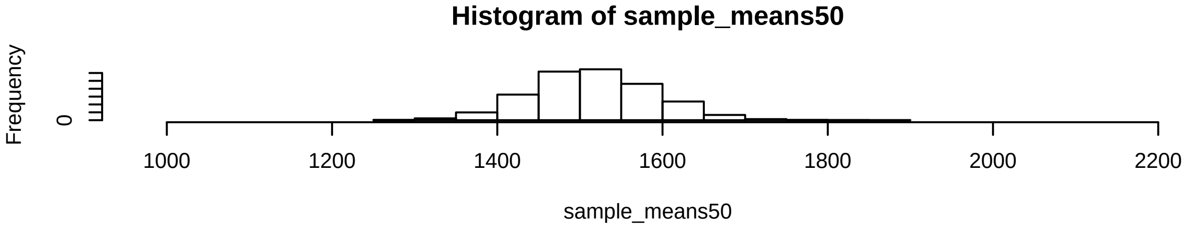
```
sample_means10 = rep(NA, 5000)
sample_means100 = rep(NA, 5000)

for(i in 1:5000){
  samp <- sample(area, 10)
  sample_means10[i] = mean(samp)
  samp <- sample(area, 100)
  sample_means100[i] = mean(samp)
}
```

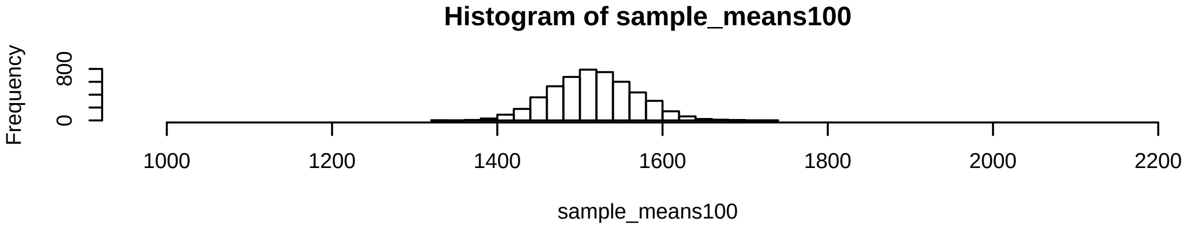
To see the effect that different sample sizes have on the sampling distribution, plot the three distributions on top of one another.

```
par(mfrow = c(3, 1))

xlimits = range(sample_means10)
```

```r
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, what happens to the center? What about the spread?

When the sample size is larger, the distribution is more of a perfect normal distribution. And the center point(mean) is more accurate. Also most of the sample mean does not diverge from the mean.

Exercises

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

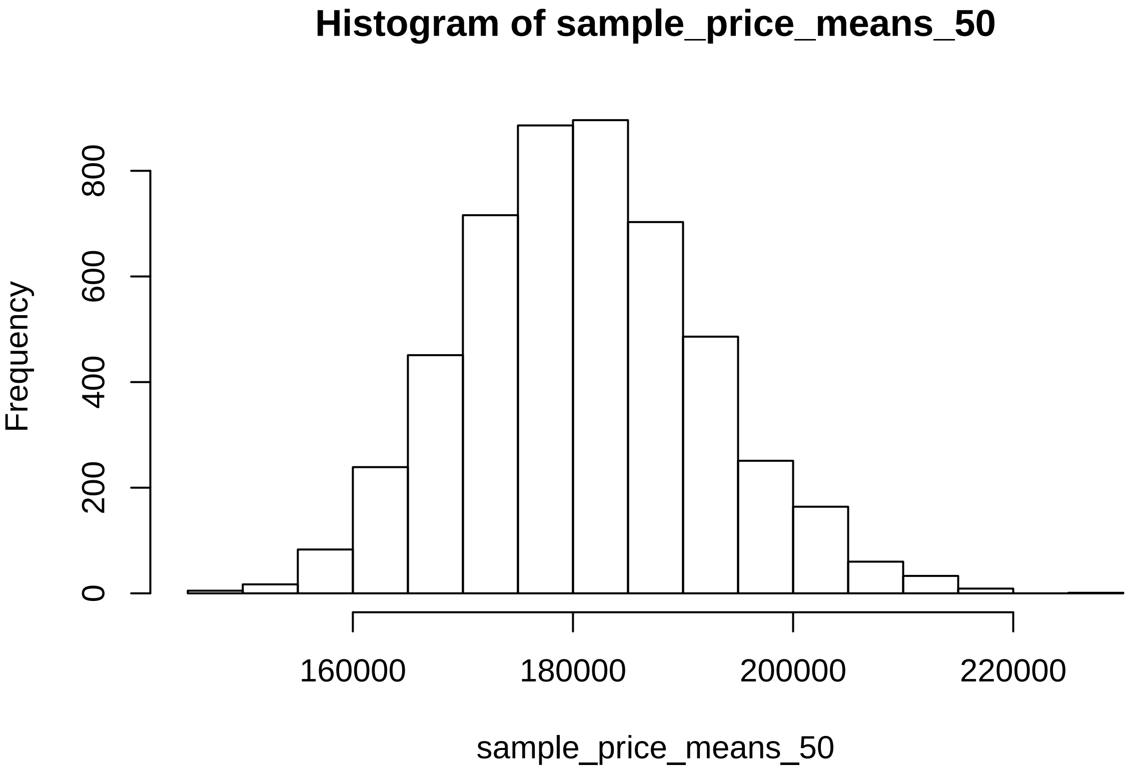
```r
set.seed(7340)

price = sample(ames$SalePrice,50)

mean(price)
```

```
## [1] 176084
```

2. Since you have access to the population, simulate the sampling distribution for 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? Finally, calculate and report the population mean.

```
sample_price_means_50 = rep(NA,5000)

for(i in 1:5000) {
  sample_price_means_50[i] = mean(sample(ames$SalePrice,50))

}

par(mfrow = c(1, 1))

hist(sample_price_means_50)
```



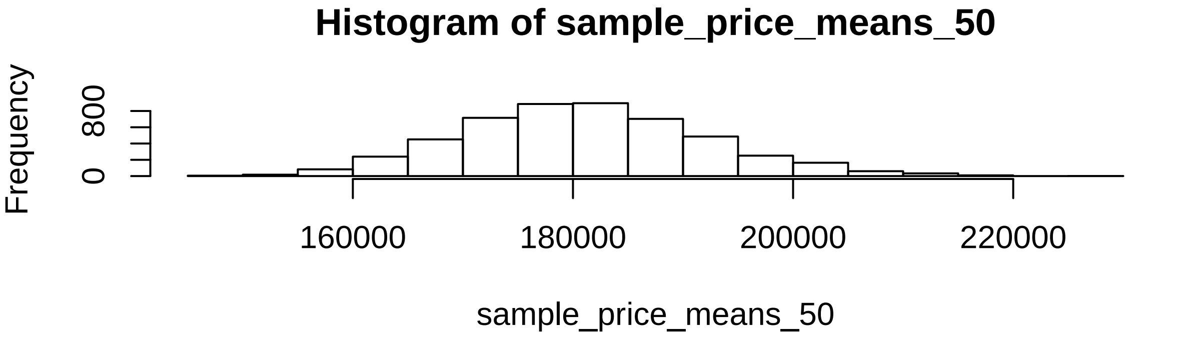
```
mean(sample_price_means_50)
```

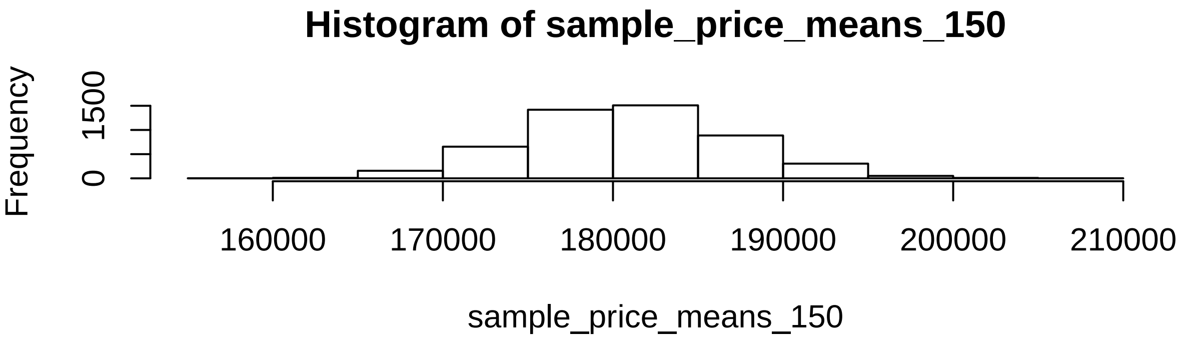
```
## [1] 181010.5
```

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?

```
sample_price_means_150 = rep(NA,5000)

for(i in 1:5000) {
  sample_price_means_150[i] = mean(sample(ames$SalePrice,150))

}

par(mfrow = c(2, 1))
```

```r
hist(sample_price_means_50)
hist(sample_price_means_150)
```





```r
mean(sample_price_means_50)
```

```
## [1] 181010.5
```

```r
mean(sample_price_means_150)
```

```
## [1] 180891.2
```

- The mean of the sample size 150 is more concentrated near the mean. It also has the perfect mean compared to the sample size of 50.

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?

- The sample size 150 is smaller spread compared to sample size 50. I would prefer the sample size of 150(Smaller spread) to make estimates.