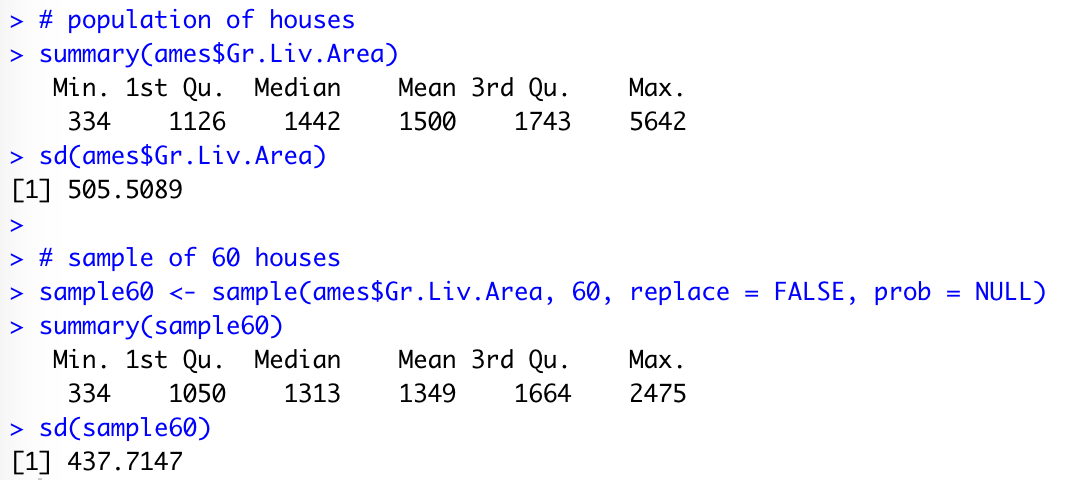
Class 5 - Lab

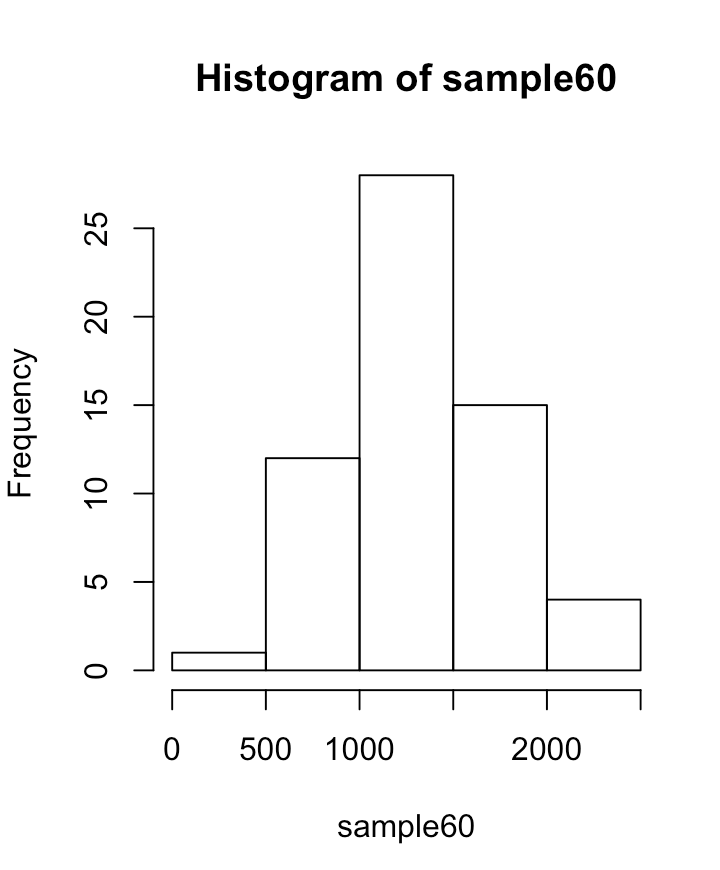
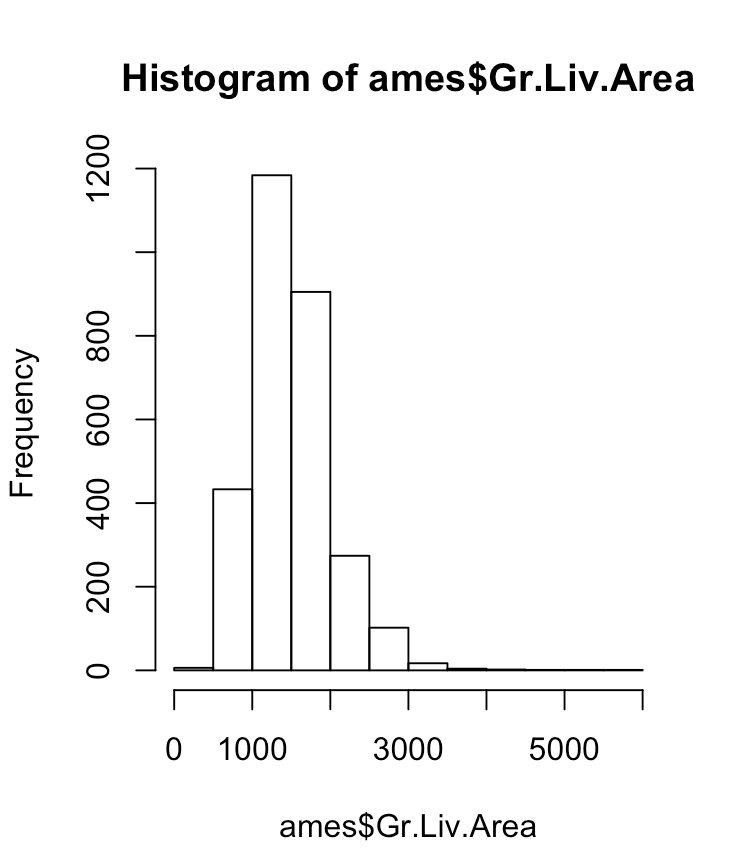
If you have access to data on an entire population, say the size of every house in Ames, Iowa, it’s straight forward to answer questions like, “How big is the typical house in Ames?” and “How much variation is there in sizes of houses?”.

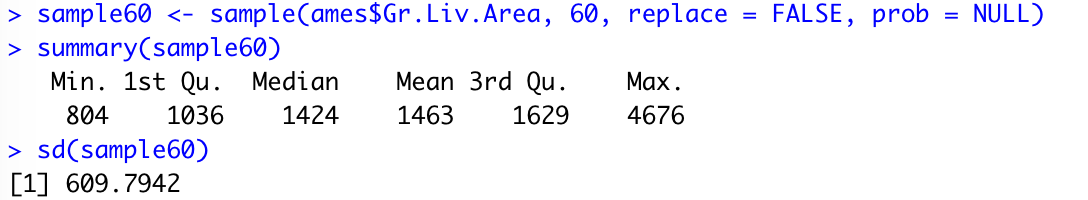
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")

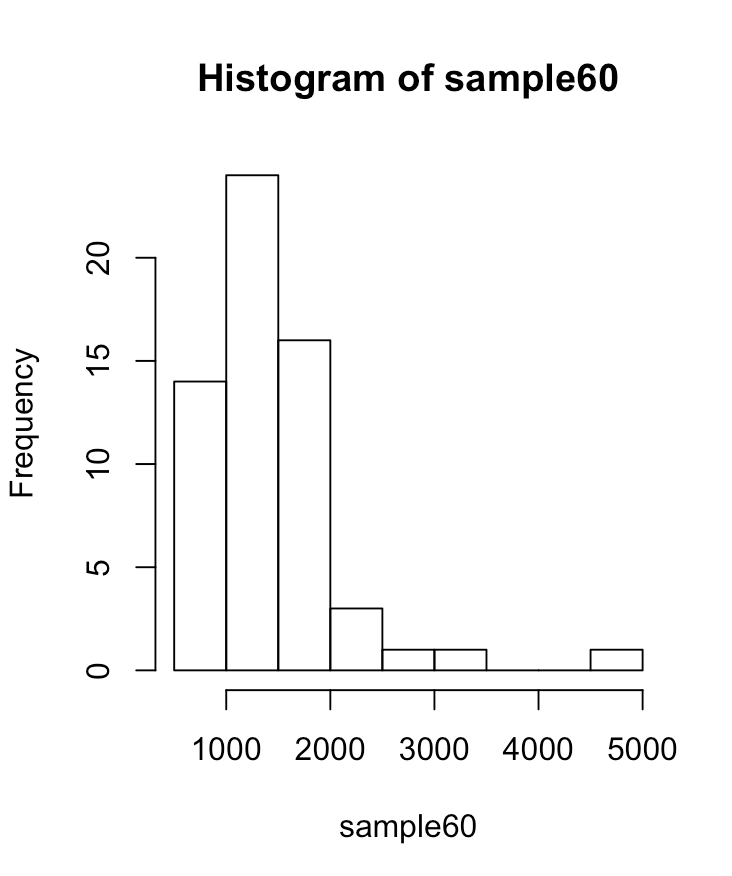
* Create a sample of 60 houses, describe the distribution of your sample and compare it to the overall population. What would you say is the “typical” size within your sample / the overall population? Also state precisely what you interpreted “typical” to mean.

The “typical,” or average size within the overall population is 1500. The average size for my sample is 1349. Since it is a random sample, the sample mean will be different every time I run the code, but will be somewhat close to the population mean of 1500.





* Create a second sample and describe and compare this one.

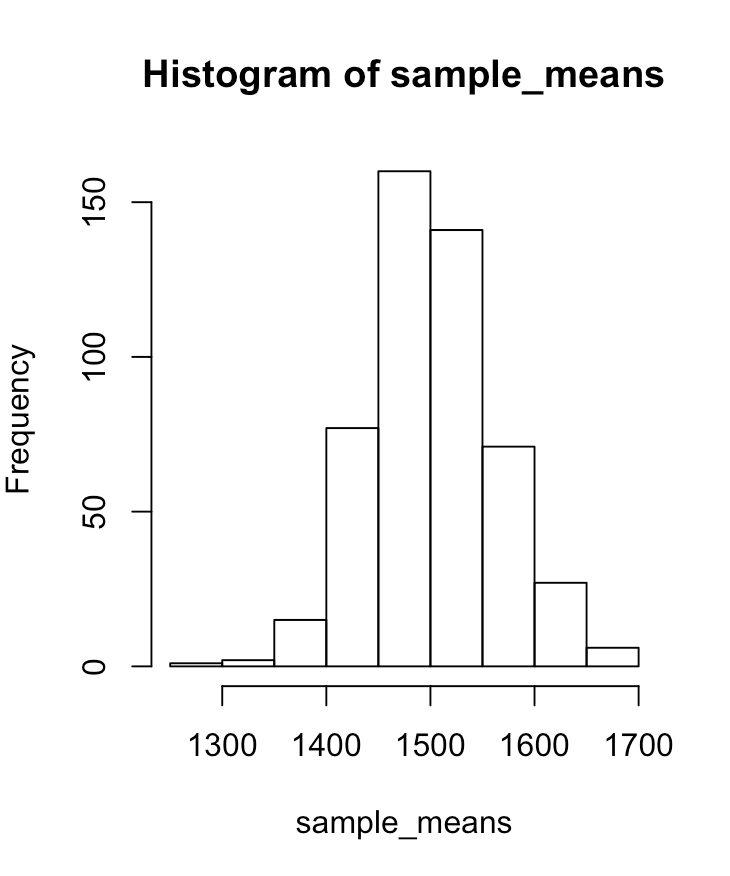
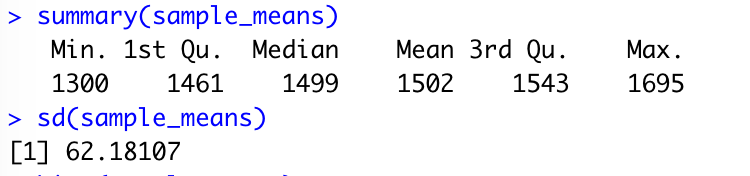


* What would be the result to take more samples? What were the results of taking larger samples? Which one would provide the more accurate estimate of the population mean?

More samples would result in a more bell-shaped curve, since the distribution would be more centered around the population mean. The Central Limit Theorem states that as the number of samples increases, the distribution of their means tends to normal. A larger sample would provide a more accurate estimate of the population mean.

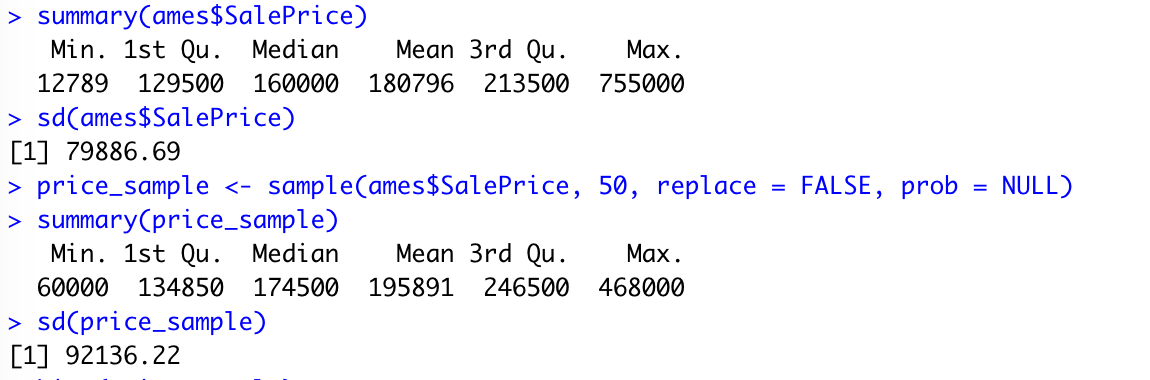
* Create a dataset with 500 independent sample means from the overall population; how does the mean distribute? What does the mean of this distribution tell us?

The mean of this distribution tells us that the sample means are centered around 1502, which is a pretty good estimate for the population mean. The distribution of the sample means is approximately normal, as per the central Limit Theorem.



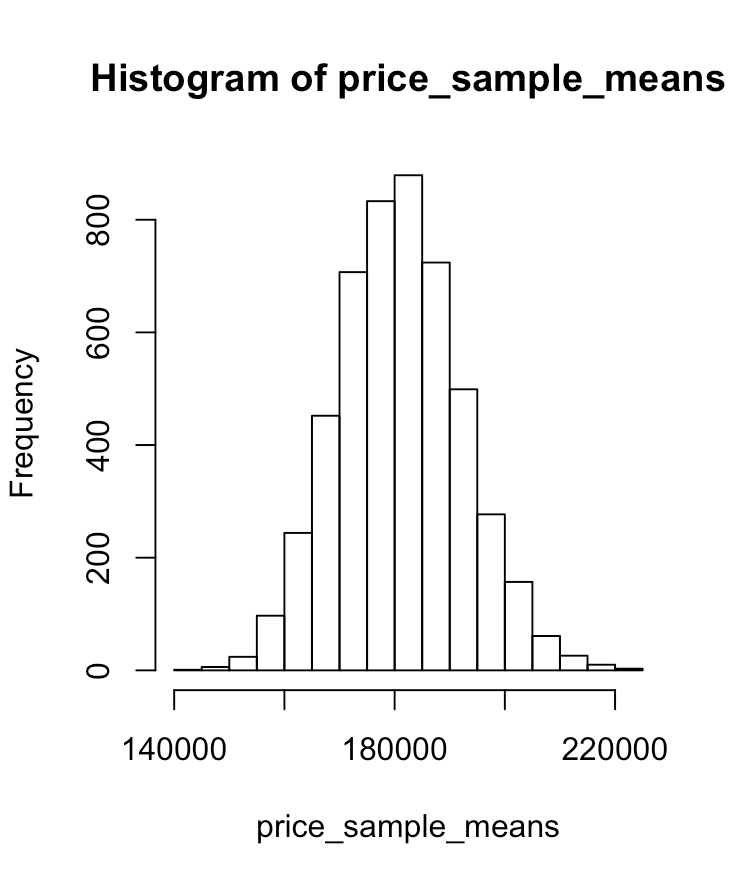
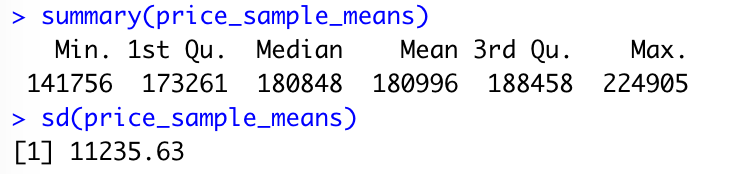
* Take a random sample of 50 from price. What is the best point estimate of the population mean?

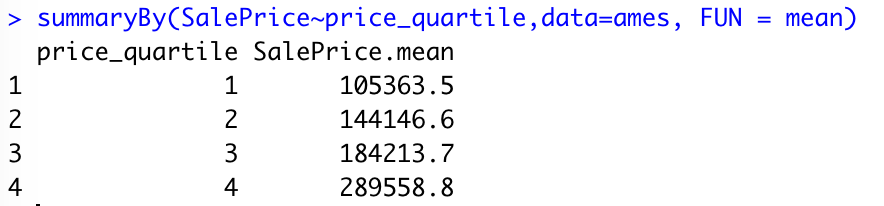
195891



* Create a dataframe with 5000 independent sample means and compare the distribution and the mean of the means to those of the overall population.

The mean is pretty close to the population mean, and the distribution is approximately normal since n is so large.



* Stratify the overall population by price (create an additional variable) and compare the sample means in the most expensive quartile and the least expensive quartile to the overall population mean).

Population mean = 180796.1

The difference between the most expensive quartile and the mean is 108762.7, and the difference between the mean and the least expensive quartile is 75432.6. The most expensive quartile is farther from the mean.