Zi Xuan Li

Professor Vahdani

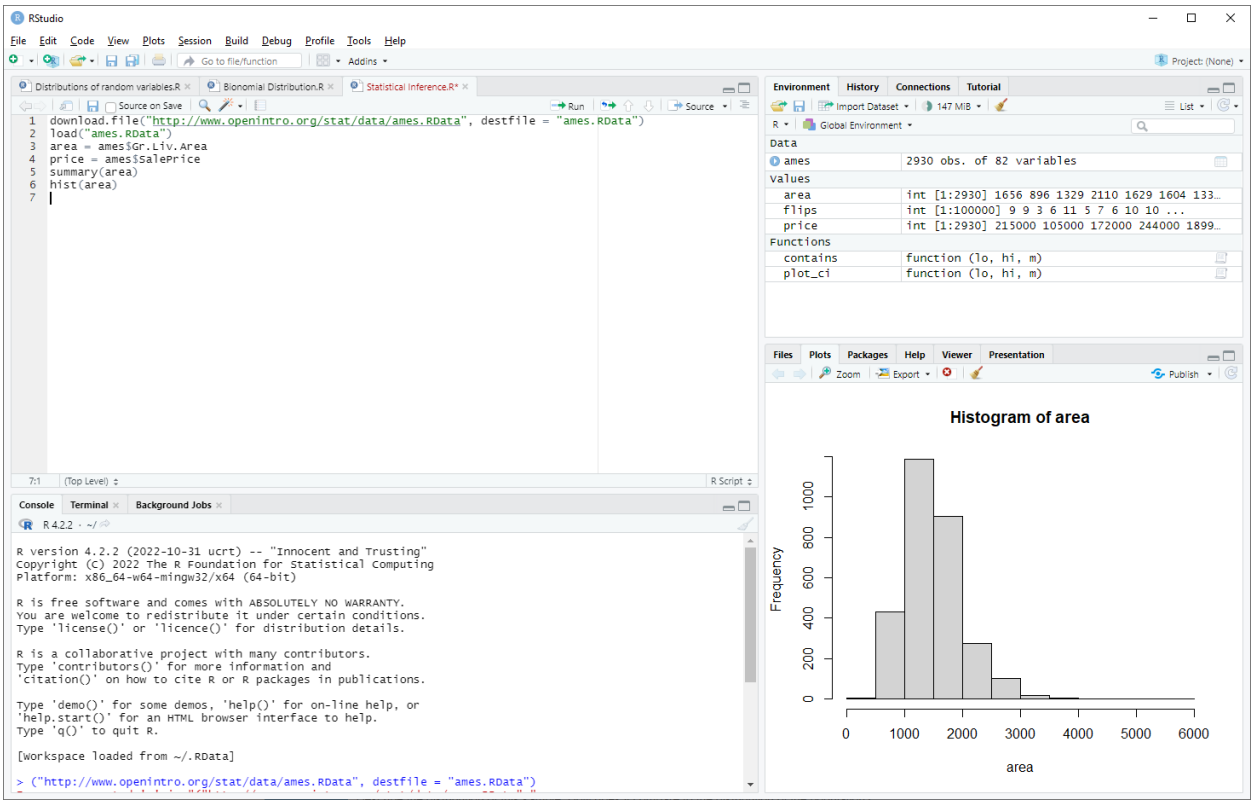
CSC 21700

December 18th, 2022

R Assignment #3

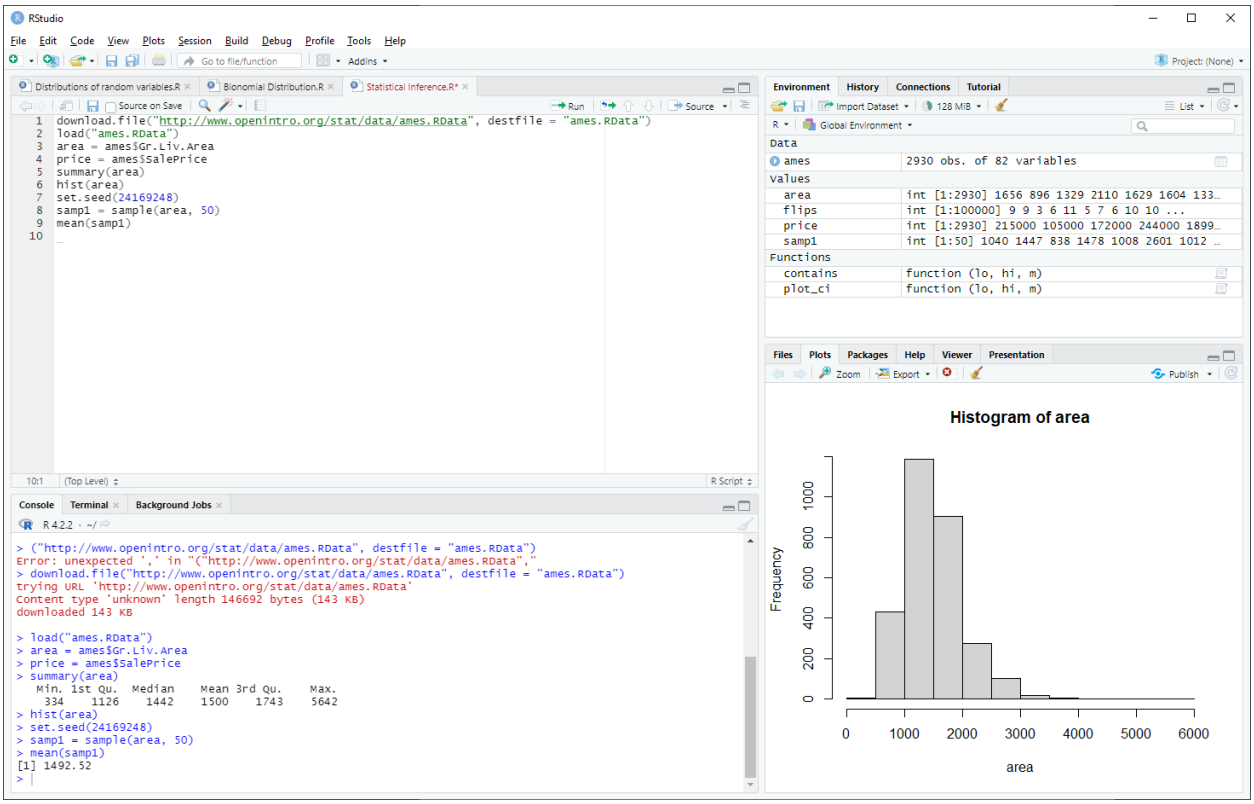
Exercise 1:

Describe this population distribution. Be sure to include a visualization in your answer.



From observing the histogram, the distribution of the living area of homes that were sold in Ames, Iowa between years 2006 and 2010 is right skewed.

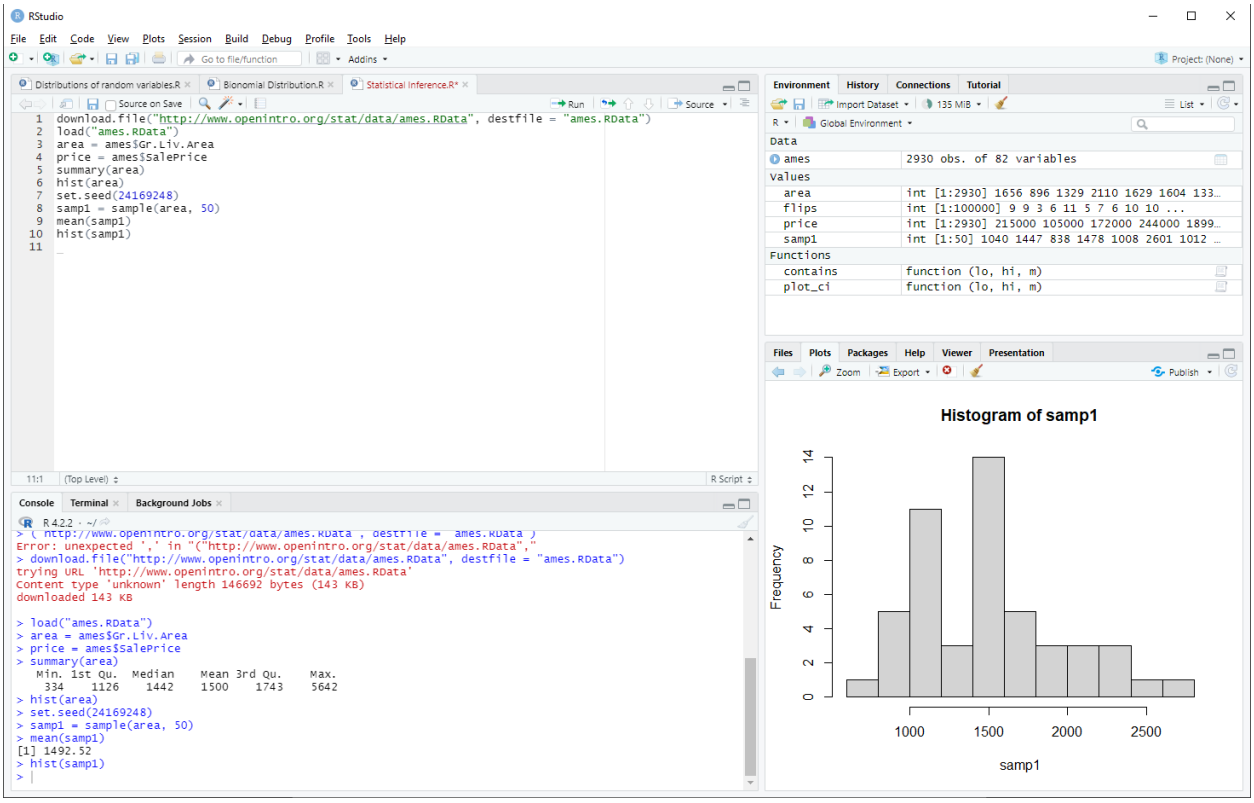
Exercise 2:



For this exercise, I was supposed to set a “random” seed so that each time we took data from the document, the random sample we obtain would be the same. The seed that I used was my employer ID number which is 24169248.

Exercise 3:

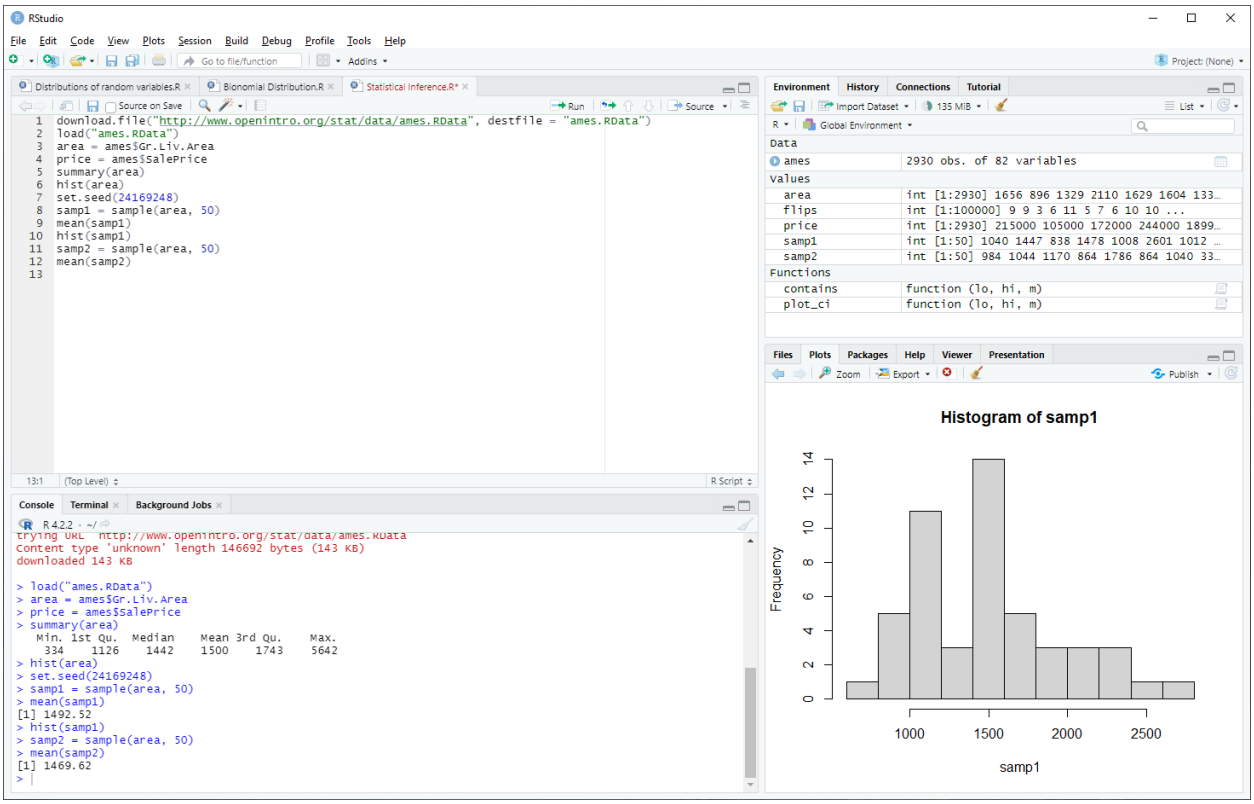
Describe the distribution of this sample? How does it compare to the distribution of the population? Be sure to include a visualization in your answer.



Based on the histogram, the distribution of the living area with the seed set to 24169248 is still right skewed. However, the range of the distribution has changed from 5,000 from the population distribution to 3,000 in the new distribution. There are also far fewer extreme outliers in the new distribution. Lastly, the new distribution is no unimodal like the initial population distribution but instead bimodal.

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? 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?

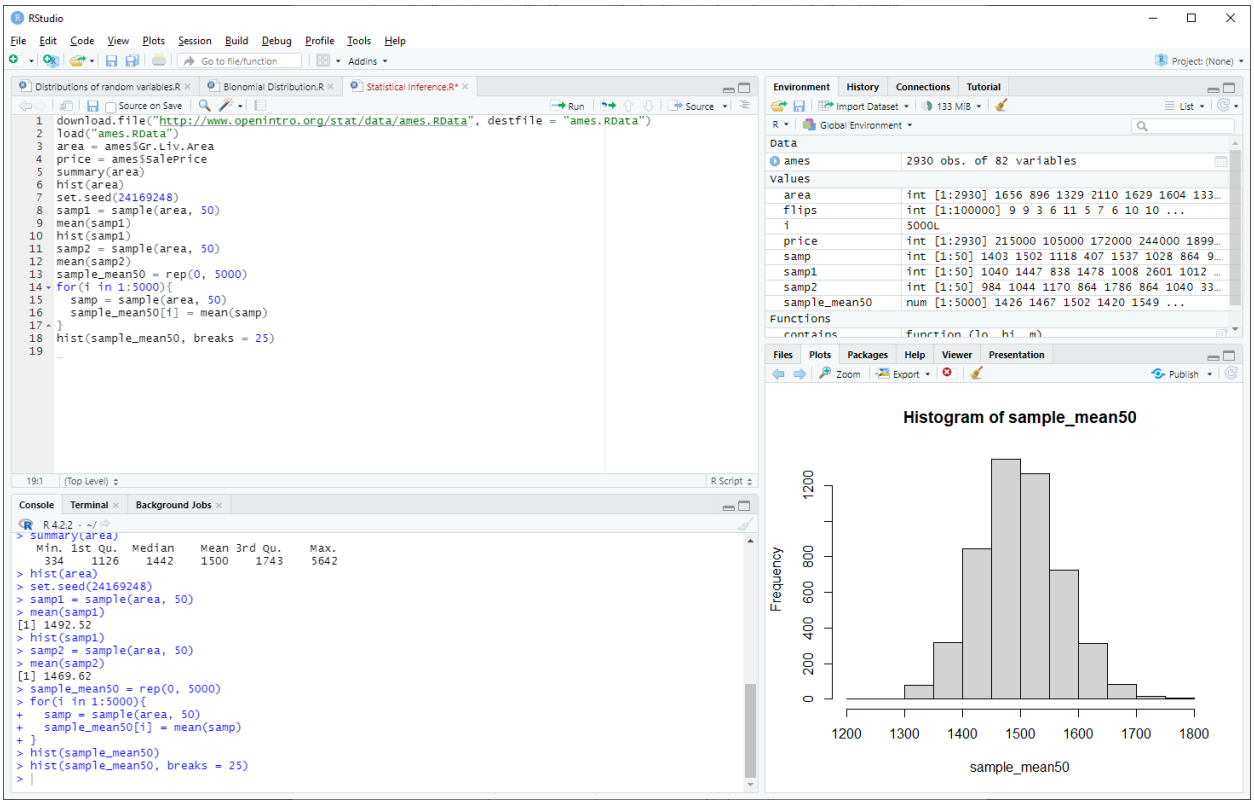


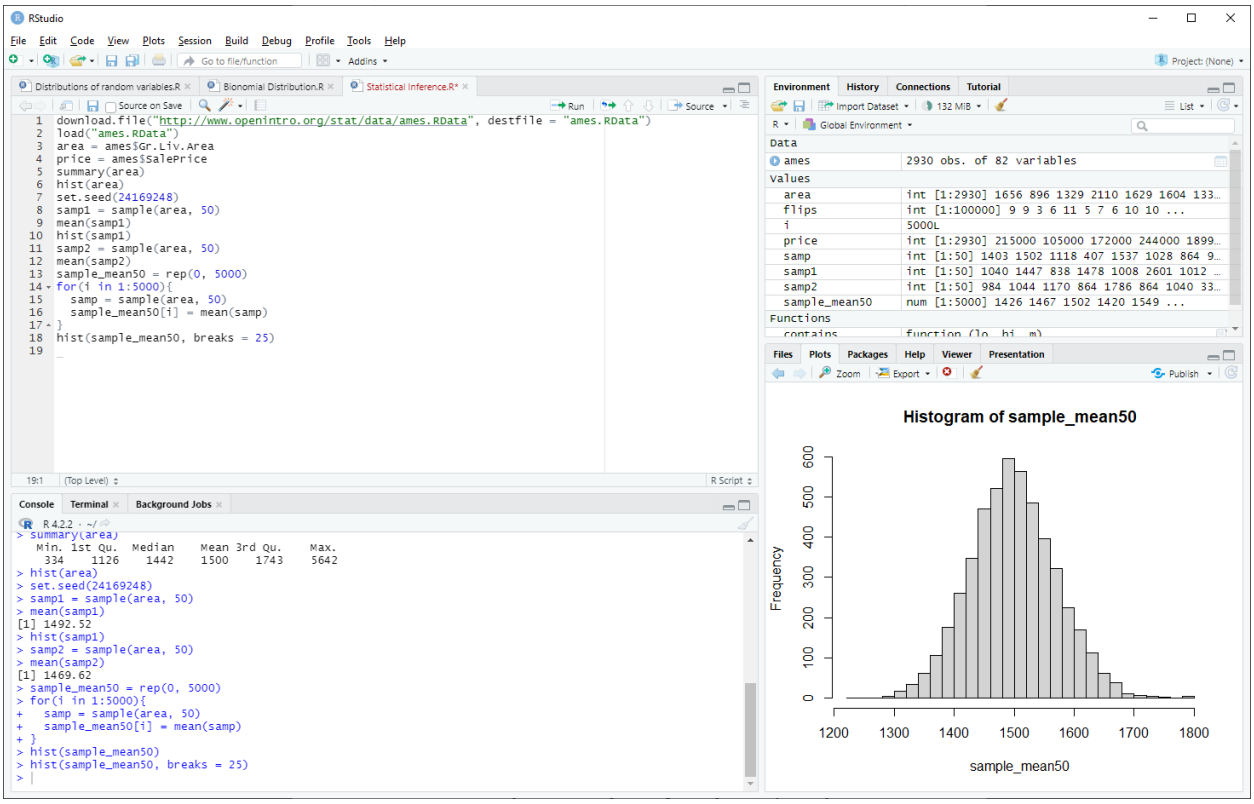
The mean of samp1 was 1,492.52 while samp2 was 1,469.62. The means of both samples are close to the true mean 1,500 but are still visibly different from each other. Using my set seed, samp1 was far closer to the true mean than samp2.

If we took a sample size of 100 and another with the size of 1,000, I would infer that the 1,000-size sample will be a more accurate estimate of the mean than the 100-size sample because larger samples have less variability than smaller samples.

Exercise 5:

How many elements are there in sample_mean50? 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?



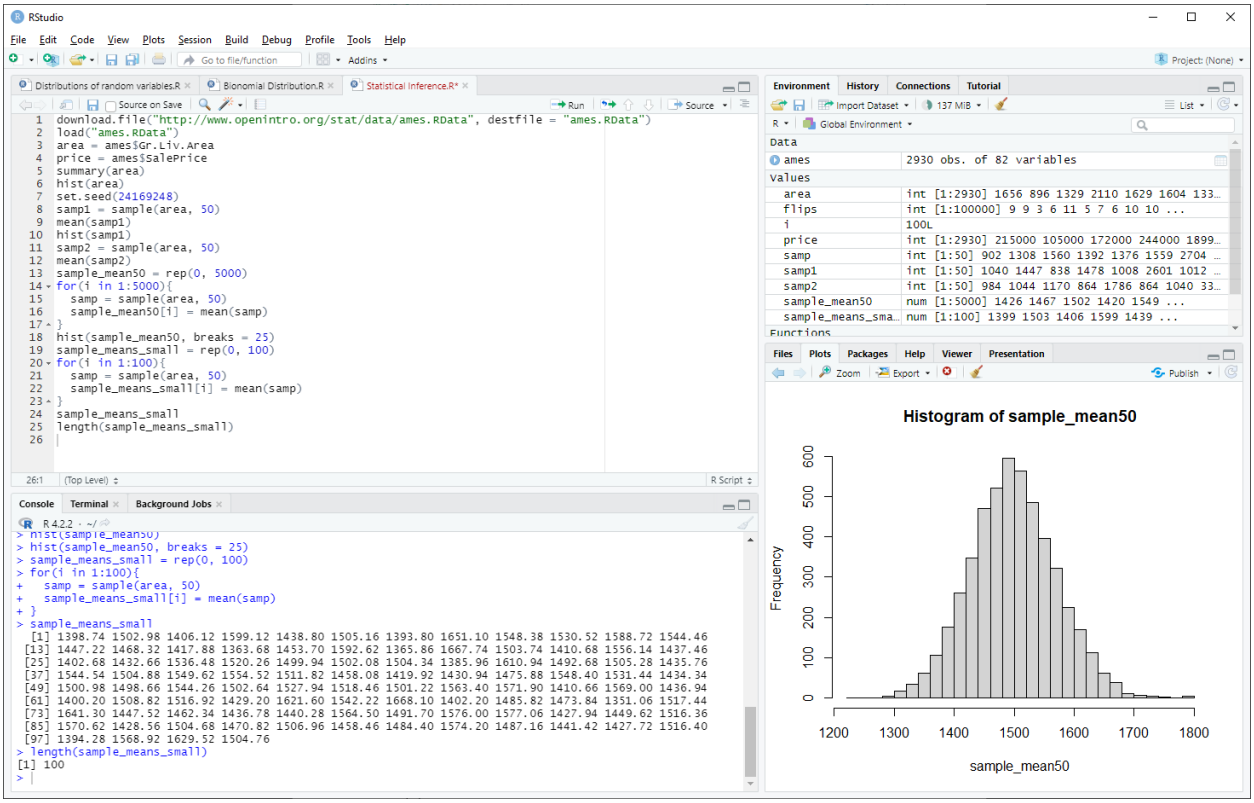


There are a total of 5,000 elements in sample_mean50. The sampling distribution looks perfectly normal, and its center is close to the true mean of 1500.

If we took the distribution of 50,000 sample means, the distribution would not deviate much from the distribution from 5,000 sample means. This is because, the distribution with 5,000 sample means is already very close to normal and anymore samples will just strengthen its normality.

Exercise 6:

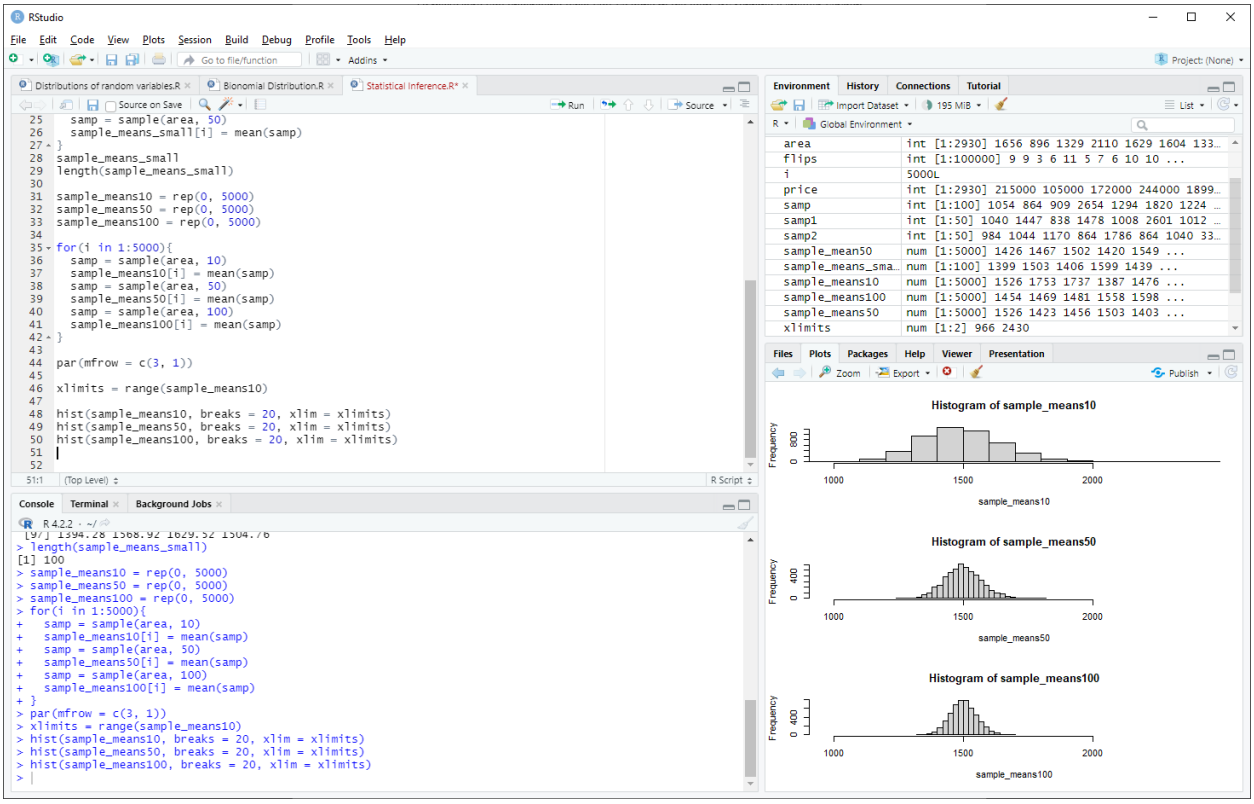
To make sure you understand what you've done in the loop, try running a smaller version. Initialize a vector of 100 zeros called sample_means_small. Run a loop that takes a sample of size 50 from area and stores the sample mean in sample_means_small, but only iterate from 1 to 100. Print the output to 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?



There are 100 elements in the object titled sample_means_small. Each element represents an individual sample mean.

Exercise 7:

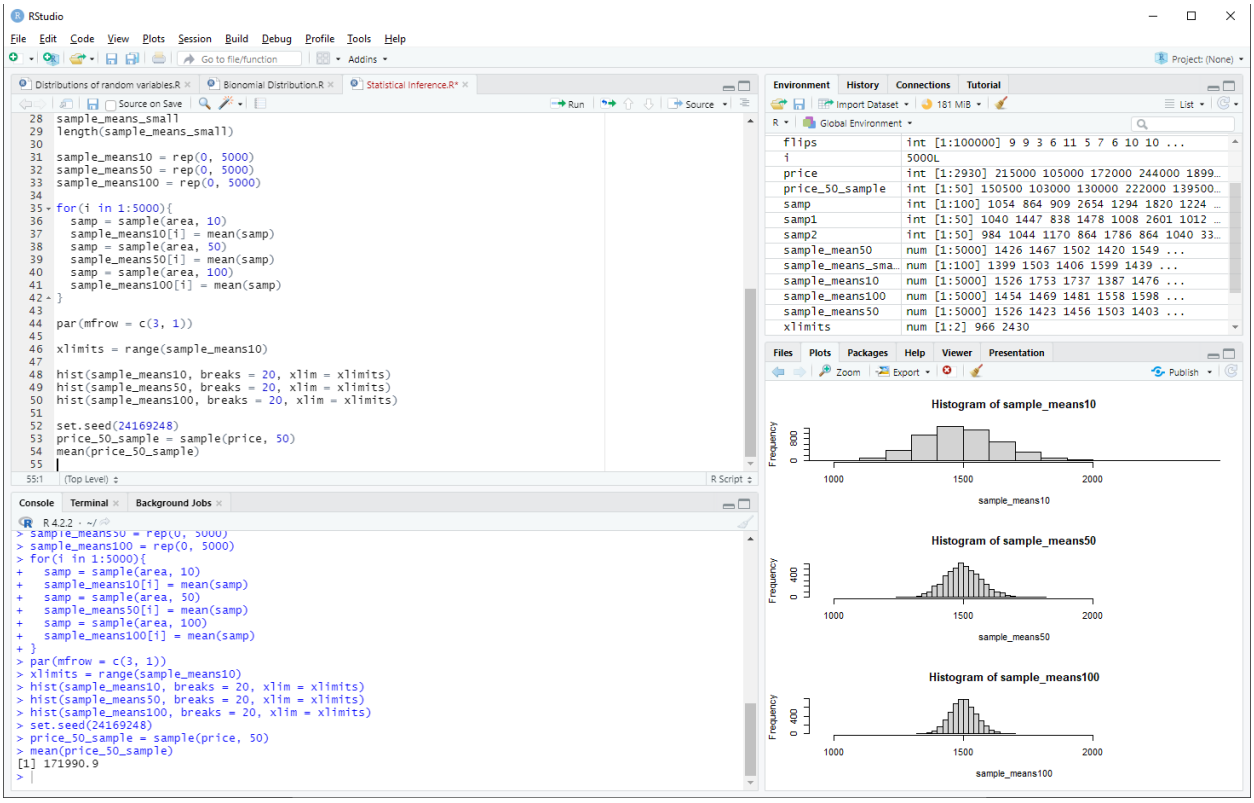
When the sample size is larger, what happens to the center? What about the spread?



Observing the 3 histograms generated from different sample means, I notice that the larger the sample size, the center moves closer to the true mean, and spread decreases.

Exercise 8:

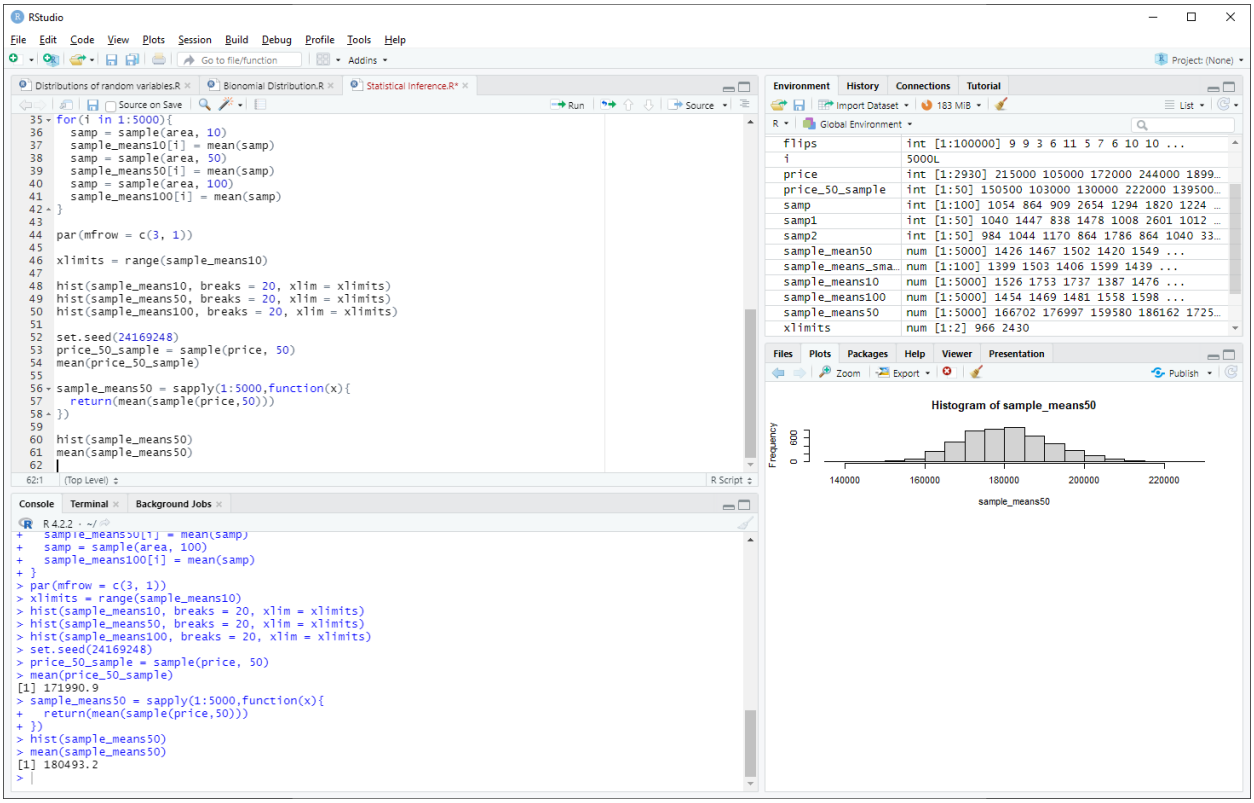
Take a random sample of size 50 from price. Using this sample, what is your best point estimate of the population mean?



The best point estimate of the population mean is the sample mean, in my case with the seed 24169248 the sample mean was 171,990.9.

Exercise 9:

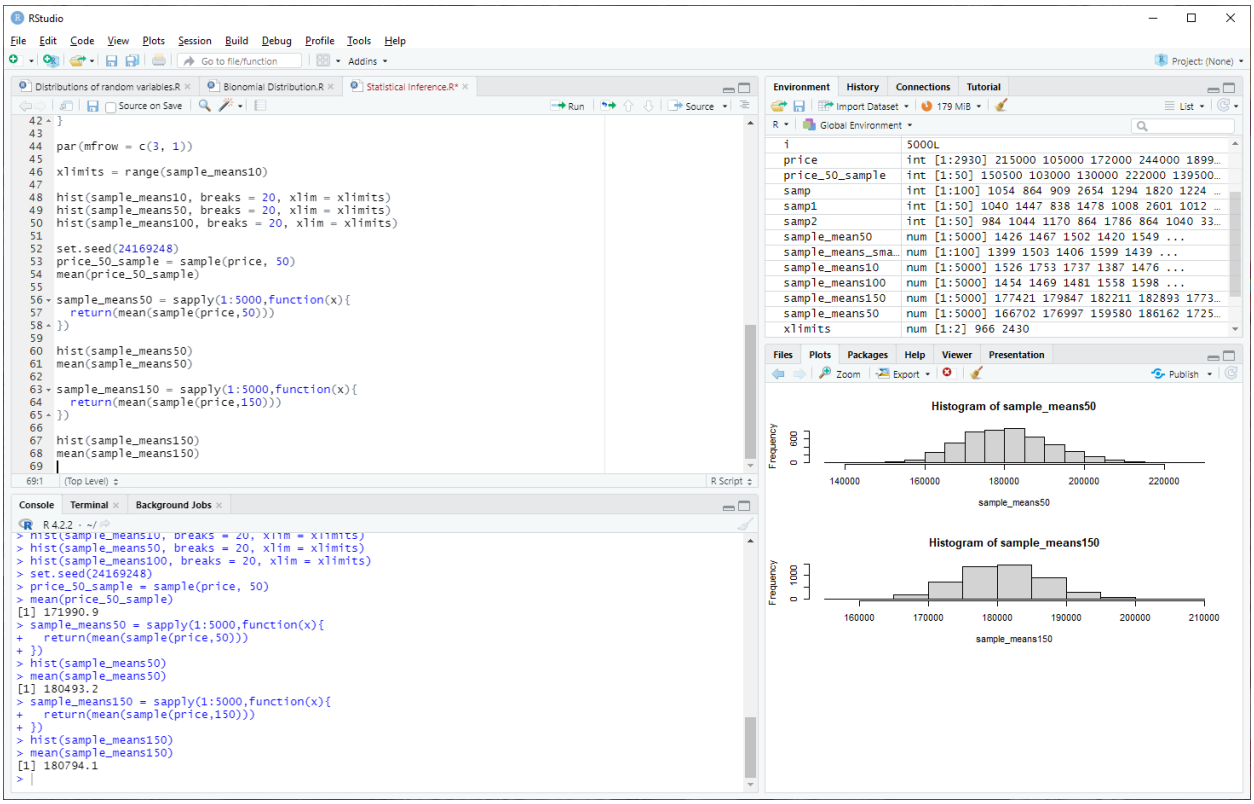
Since you have access to the population, simulate the sampling distribution for x price by taking 5000 samples from the population of size 50 and computing 5000 samples means. Store these means in a vector called sample_means50. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean home price of the population to be? Finally, calculate and report the population mean.



Based on the histogram, the sampling distribution looks normal, and I would assume the mean home price is around 180,000. The exact calculation for the population mean is 180493.2.

Exercise 10:

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. Plot the data, then 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?



Based on the histogram, the shape of this sampling distribution is normal and is centered around 180,000. Comparing it to the sample size 50 distribution, sample size 150 distributions has far less spread. I would still assume that the mean sale price of the homes in Ames is still 180,000.

Exercise 11:

Of the sampling distributions from 9 and 10, 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 from 9 and 10, the sampling distribution from exercise 10 has smaller spread. If we are looking to make estimates that are often closer to the true value, we would prefer a distribution with small spread because there is far less uncertainty about the range of values of the mean when the spread is less.