Lab 4

**1) Describe this population distribution.**

The histogram for area is right skewed, with most data on the left side of the graph. It is unimodal in shape and not very symmetric. The center of the histogram is around 1500 sq ft, which is the mean. It spans a range of 334 to 5643 sq ft. 25th percentile of people live in a home up to 1126 sq ft, and 75th percentile live in a home up to 1743 sq ft.

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

Summary(samp1)

Hist(samp1)

The distribution of this sample of 50 homes is close the overall area vector, but with a smaller range. Both are unimodal in shape and right skewed. The quartiles, medians, and means are very close, and the histogram has a peak around 1400-1500 sq ft. It is not symmetric as well. The 1st quartile is slightly smaller than the previous sample, but the 3rd quartile is slightly bigger. The sample gives a good picture of

**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?**

mean(samp1) = 1536.64

samp2 <- sample(area, 50)

mean(samp2) = 1620.64

The mean of samp2 is bigger than the mean of samp1 by about 84 sq ft.

The sample of size 1000 would provide a more accurate estimate of the population mean because it is much closer to the overall size of the population, so there will be less variability and error. Outliers will be less likely to skew the average data with a larger sample size.

**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?**

There are 5,000 collected sample means. The histogram shows a much more normal distribution, with a unimodal, symmetric shape and a peak between 1480-1500 sq ft. If we instead collected 50,000 sample means, the distribution would not change too much, but it would be much more optimized to show the accurate mean of the overall population, as there is less standard error with a higher sample size.

**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. 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)

}

hist(sample\_means\_small)

There are 100 elements in the object sample\_means\_small, and each element represents one mean of a sample size of 50 randomly selected house areas from the entire population of the dataset.

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

The spread narrows when the sample size is larger, and the center moves closer to the population’s center distribution to give a more accurate simulation of the population. The peaks do not seem to move too much, but they do move a little to hover closer to around 1500.

**ON YOUR OWN**

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

sampp <- sample(price, 50)

summary(sampp)

Min. 1st Qu. Median Mean 3rd Qu. Max.

44000 128000 159500 182118 215283 462000

The mean is around $182,118 based on the sample of 50 house prices generated randomly.

1. **Since you have access to the population, simulate the sampling distribution for x¯pricex¯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, 50)

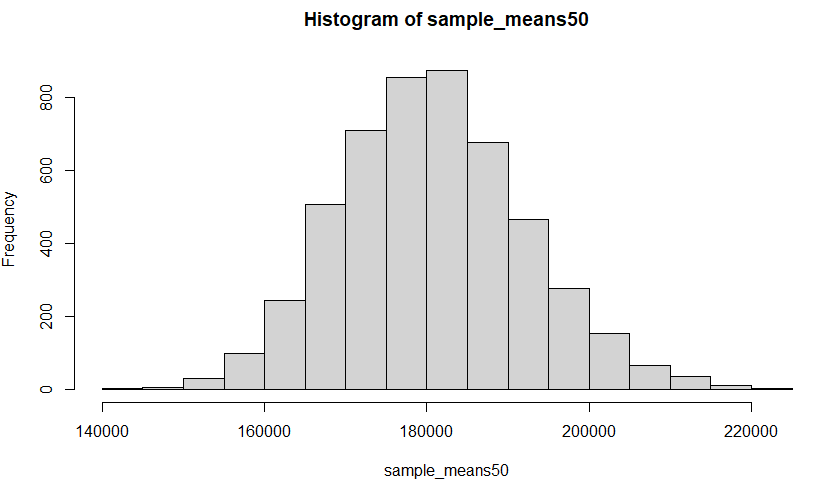
sample\_means50[i] <- mean(samp)

}

summary(sample\_means50)

Min. 1st Qu. Median Mean 3rd Qu. Max.

140866 172796 180311 180718 187935 224608

hist(sample\_means50)

Based on the for loop calculation of 5000 samples of 50 randomly selected house prices, the mean is $180,718. The shape of the histogram is a normal distribution with a unimodal, symmetric shape. The peak appears to be slightly over $180,000, which lines up with the summary table.

mean(price)= 180796.1

The actual mean is $180,796.10 and is very close to the simulated samples’ mean.

1. **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){

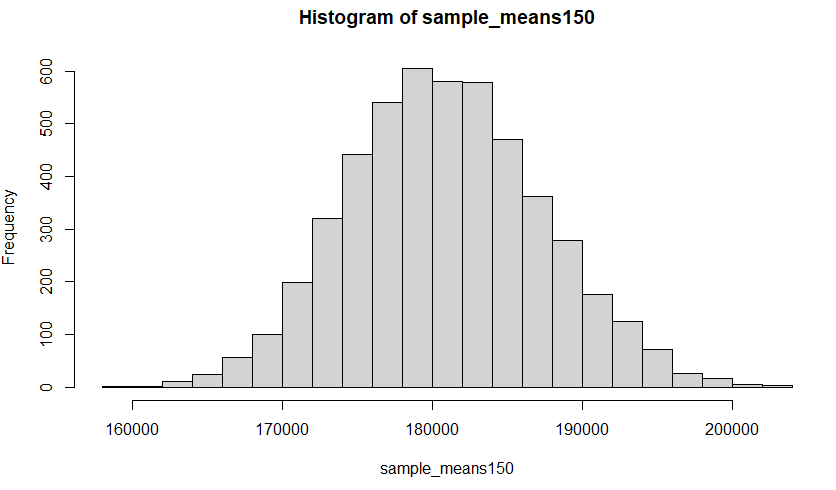
samp <- sample(price, 150)

sample\_means150[i] <- mean(samp)

}

summary(sample\_means150)

Min. 1st Qu. Median Mean 3rd Qu. Max.

 158405 176444 180667 180881 185178 203537

hist(sample\_means150, breaks = 25)

The histogram for sample size 150 also has a normal

distribution and a unimodal, symmetric shape. Its peak

on the graph appears a little under $180,000, but that

may be due to the smaller spread of the graph due to

the larger sample size. The mean based on this is ~$180,881.

1. **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 sampling distribution from 2 has a smaller spread because it has a smaller sample size for each element (50 vs. 150). As shown on the histograms, the range seen on #2’s histogram is $140,000 - $220,000, while for #3, it is $160,000 - $200,000. We would prefer a smaller spread to make an estimate closer to the true value, because this means there is less standard error due to having a larger sample size.