SEIS 631 Foundations of Data Analysis
Assignment 4

In this assignment, we 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.

The data: We consider 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. Our particular focus for this lab will be all residential home sales in Ames between 2006 and 2010. This collection represents our population of interest (which is rare to have access to), but we will also work with smaller samples from this population.

Load the data provided in the ames.RData file. (See previous assignments if you forget how.)

We see that there are quite a few variables in the data set, enough to do a very in-depth analysis. For this assignment, 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**). To save some effort throughout the assignment, create two variables with short names that represent these two variables.

```
area <- ames$Gr.Liv.Area
price <- ames$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)
hist(area)
```

Question 1 [MULTIPLE CHOICE] **Which of the following is *false*?**
(a) The distribution of areas of houses in Ames is unimodal and right-skewed.
(b) 50% of houses in Ames are smaller than 1,442 square feet.
(c) The middle 50% of the houses range between approximately 1,130 square feet and 1,740 square feet.
(d) The IQR is approximately 610 square feet.
(e) The smallest house is 534 square feet and the largest is 3,642 square feet.

The unknown sampling distribution: In this assignment, 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 sample function to sample from the population.

```
samp0 <- sample(area, 50)
```

This command collects a simple random sample of size 50 from the vector **area**, which is assigned to **samp0**. This is like going into the City Assessor's database and pulling up the files on 50 random home sales. If we didn't have access to the population data, working with these 50 files would be considerably simpler than having to go through all 2930 home sales. Now that you've taken a sample, take another sample and compare the two. Create histograms and summaries of each.

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

Question 2: Describe the distribution of this sample? How does it compare to the distribution of the population?

> ***GGplot Extension (optional)***: We can compare the two samples a little more directly using a density plot (**geom_density()**) from **ggplot**. A density plot is like the frequency polygon (**geom_freqpoly()**) but it smooths the

curve a little. This is useful for smaller samples because it can help to eliminate some of the noise in the data. First we need store our samples in a data frame:

```
samp.df <- data.frame(samp0 = samp0, samp1 = samp1)
```

Now we can build the **ggplot** graph. (Remember to load the **ggplot2** library.)

```
library(ggplot2)

ggplot(data = samp.df) +
    geom_density(aes(x = samp0, color = "Sample 0")) +
    geom_density(aes(x = samp1, color = "Sample 1"))
```

Notice that you can add two plots to the same graph just by adding another layer. Now we can compare the two samples directly.

If 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)
```

Depending on which 50 homes you selected, your estimate could be a bit above or a bit below the true population mean of approximately 1,500 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 less than 2% of the population.

Question 3 [MULTIPLE CHOICE] **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?**
(a) Sample size of 50
(b) Sample size of 100
(c) Sample size of 1000

Not surprisingly, every time we take another random sample, we get a different sample mean. 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. In this assignment, because we have access to the population, we can build up the sampling distribution for the sample mean by repeating the above steps many times. Here we will generate 5000 samples and compute the sample mean of each.

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

Remember, 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.

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

Above, we use R to 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, using a *for loop*.

Challenge: *Recreate these graphs using* **ggplot2**.

Question 4: Describe the sampling distribution (the distribution of the sample means that you just created), and be sure to specifically note its mean.

Question 5: [MULTIPLE CHOICE] **Which of the following is *true* about the elements in the sampling distributions you created?**
(a) Each element represents a mean square footage from a simple random sample of 50 houses.
(b) Each element represents the square footage of a house.
(c) Each element represents the true population mean of square footage of houses.

The sampling distribution that we computed 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 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 sampling 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)
hist(sample_means10, breaks = 20, xlim = xlimits)
hist(sample_means50, breaks = 20, xlim = xlimits)
hist(sample_means100, breaks = 20, xlim = xlimits)
```

The first command specifies that you'd like to divide the plotting area into 3 rows and 1 column of plots. The **breaks** argument specifies the number of bins used in constructing the histogram. The **xlim** argument specifies the range of the x-axis of the histogram, and by setting it equal to **xlimits** for each histogram, we ensure that all three histograms will be plotted with the same limits on the x-axis.

> ***GGplot Extension (Optional):*** We can also do this with **ggplot2** like this:
>
> ```
> samp.df2 <- data.frame(samp.size = rep(c(10,50,100),each = 5000),
>                        samp.means = c(samp_means10, samp_means50, samp_means100))
>
> ggplot(samp.df2, aes(x = samp.means)) +
>        face_grid(rows = vars(samp.size)) +
>        geom_histogram(col = "black", alpha = 0.2)
> ```
>
> Alternatively, we could use density plots like we did earlier (using the samp.df2 created in the code block directly above).

```
ggplot(samp.df2, aes(x = samp.means, color = as.factor(samp.size))) +
    geom_density()
```

In both of these plots, we first needed to add a variable (samp.size) which recorded the sample size for each simulated sample. This is a grouping variable, and grouping variables are often called "factors." GGplot can easily separate or group things by "factor" variables. In this case, however, because the sample sizes are all numbers, R by default, treats them as numeric variables. The force it to treat sample size as a grouping variable we use the **as.factor()** function.

Question 6: It makes intuitive sense that as the 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, what can you say about the variability of the sampling distribution? Explain your answer.

So far, we have only focused on estimating the mean living area in homes in Ames. Now you'll try to estimate the mean home price.

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

Question 8: Since you have access to the population, simulate the sampling distribution of 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?

Question 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?

Question 10: [MULTIPLE CHOICE] Which of the following is false?
(a) The variability of the sampling distribution with the smaller sample size (sample means50) is smaller than the variability of the sampling distribution with the larger sample size (sample means150).
(b) The means for the two sampling distributions are roughly similar.
(c) Both sampling distributions are symmetric.

Submission: The submit your answers to the 10 questions via Canvas. Save your code blocks and output in case you have questions or concerns in the future.