For this project, you will generate several sampling distributions of the sample mean with Fathom 2 and use these distributions to confirm the Central Limit Theorem. You will be sampling from a population of 686 randomly selected pennies. The variable of interest is the age of the pennies, which you will find has a decidedly non-normal distribution. You may work alone or with a partner, but be aware that you will receive the same grade as your partner. I will be assessing your mastery of Fathom 2 based on what you submit to me, so take this project seriously. Your grade for this project will count as 5% of your final course grade. Your project report will be due at the start of class on October 28.

1. Begin by downloading the Fathom 2 data collection from either Blackboard or the Rowan openarea (see the syllabus for directions). The file is named Penny_Project08.ftm. Save the file to a USB drive or your H: drive; any files saved to the desktop of a campus cluster computer will be deleted.

2. To access Fathom 2 from a campus computer lab, log onto your Rowan network account (same as for email) and then click on the Start menu in the lower left corner of the screen. Follow this path: Network Applications > Applications > Mathematics > Fathom 2. The Fathom Help menu is very useful for getting started.

3. Open up the Penny_Project08 file. You will see a collection named Penny_Population that is comprised of the years stamped on a random sample of 686 pennies. I created a case table with a Year attribute to get you started. Scroll through the table and confirm that it consists of 686 cases.

4. Select the case table, and then select Show Formulas under the Table menu. Create a new attribute named Age. Double click on the formula line for your Age attribute. Enter the formula Age = 2008 - Year, and click on OK.

5. For the purposes of this project, we are interested in the Age attribute. Plot a histogram of the Age attribute. Select the histogram, and go to the Graph menu at the top of the screen. Click on Scale | Relative Frequency. Print your relative frequency histogram to include in your project report. This histogram depicts the relative frequency distribution of the ages of the pennies comprising our "population." You should recognize from your histogram that the population of penny ages does not have a normal distribution. Instead, your histogram provides a good example of a geometric distribution. Practically speaking, why might the distribution of penny ages take this form?

6. Create a summary table, and use it to compute the population mean (μ) and standard deviation (σ) for the Age attribute. You will need to add the formula PopStdDev() to the summary table. Include a printout of this summary table in your report.

7. Now that you are familiar with the population of penny ages, let’s build some sampling distributions. Start by taking a random sample of size n = 5. Select the Penny_Population collection and choose the Sample Cases option from the Collection menu. By default, Fathom takes a sample of ten cases with replacement and places them in a new collection named Sample of Penny_Population. You’ll need to change these settings in the inspector. Double-click on the new collection, and change the
settings on the Sample panel to match those shown to the right. Click on Sample More Cases to collect a sample of the proper size.

8. Next, go to the Measures panel on the inspector. Define the measures MeanAge and SampleSize as shown to the right. Doing so will allow you to compute the mean age for each sample of 5 pennies that you select. To simulate the sampling distribution of the sample mean for all possible samples of size \( n = 5 \), we will collect 1000 measures. In other words, 1000 samples, each of size \( n = 5 \), will be drawn from the original population. For each sample, the mean \((\bar{x})\) will be calculated, and the pennies will be returned to the population before the next sample is drawn. Note that the inspector shows the sample mean for my sample; the value shown on your inspector will almost certainly be different.

9. Double click on the Sample of Penny_Population collection. Then, select Collect Measures under the Collection menu at the top of the screen. You will need to double click on the newly created Measures from Sample of Penny_Population collection to bring up its inspector. Change the settings to match those shown to the right. Then, click on Collect More Measures. It will take some time to collect the measures. What Fathom is doing is drawing a random sample of 5 pennies, calculating the mean age for the sample, replacing the pennies back into the population, and then repeating the process until 1000 sample means have been collected. These 1000 sample means will serve as an approximation of “all possible samples” of size \( n = 5 \).

Plot a relative frequency histogram of the 1000 sample means you just generated. Open the graph inspector and change the binWidth to 1. Print a copy of your graph for your project report now, as it will be difficult to do so once you complete step 11. This histogram represents a simulated sampling distribution of the sample mean for \( n = 5 \). How does the shape of this sampling distribution compare to that of the population? Use a summary table to compute \( \mu_{\bar{x}} \) and \( \sigma_{\bar{x}} \); your variable of interest is the MeanAge attribute for the Measures from Sample of Penny_Population collection. Include a printout of this summary table in your project report. According to the Central Limit Theorem, what are the expected values for \( \mu_{\bar{x}} \) and \( \sigma_{\bar{x}} \)? How do the values you calculated compare to the expected values? What are some possible reasons for any discrepancies?

11. We are ultimately interested in comparing the sampling distributions of the sample mean for \( n = 5 \) and \( n = 50 \), so you will need to repeat steps 7-10 above with a few modifications. First, substitute 50 for 5 in the inspector for the Sample of Penny_Population collection. Also, uncheck the “Replace existing cases” box on the inspector for the Measures from Sample of Penny_Population collection. Doing so will allow you to save your measures for \( n = 5 \) and \( n = 50 \) together in a single collection. Click on Collect More Measures to generate 1000 sample means for \( n = 50 \). It will take some time to compute these sample means, so be patient with your computer. Confirm that you now have 2000 cases in a single Measures from Sample of Penny_Population collection.
12. Now, you are ready to plot a split histogram that displays both of your sampling distributions. (Hopefully you already printed the histogram requested for step 10; it is no longer easy to do so.) First, change the scale on your histogram from relative frequency to density. Doing so will allow you to superimpose the theoretical normal distributions later. Next, open the inspector for the Measures from Sample of Penny_Population collection and go to the Cases tab. Drag the SampleSize attribute to the vertical axis of the histogram for MeanAge while simultaneously holding down the Shift key. The histogram will split in half, showing the sampling distributions for both n = 5 and n = 50. Double click on your graph to open its inspector. Change the bin width to 1, and adjust the yUpper setting if necessary.

13. Superimpose the theoretical normal distributions on top of your histograms. Select the graph, and open up the Graph menu at the top of the screen. Select Plot Function, and type in this function: normalDensity(x, mean(), PopStdDev()) and click on OK. Print your split histogram to include in your project report. How well does a normal curve fit each of the sampling distributions? What is the effect of sample size on the fit?

14. Select the graph you just generated. Go to the Object menu at the top of the screen, and select Duplicate Graph. Delete the normalDensity function from the duplicate graph. Then, go to the pull down menu at the upper-right corner of the duplicate graph, and select Normal Quantile Plot. Consult section 7.7 in your textbook for a detailed description of how this plot is used for assessing normality. In short, the closer the normal quantile plot is to a straight line, the better a normal distribution “fits” your sampling distribution. What is the effect of sample size on how close a sampling distribution comes to being normal?

15. Compute the mean and population standard deviation for the MeanAge attribute of each of your sampling distributions. Note that the values have changed in the summary table you created in step 10. This happened because Fathom pooled together all the data for samples of size 5 and size 50. Separate out the two data sets by opening up the inspector for the Measures from Sample of Penny_Population collection again and dragging the SampleSize attribute to the vertical arrow on the summary table while once again holding down the shift key. You will probably need to resize your summary table in order to view all of it. Print a copy of this summary table for your project report.

16. Proceed to the next page for directions on how to write up your project report.
Writing up your Project Report

If you have completed this simulation with a partner, submit one report with both of your names on it. Your report will consist of several Fathom printouts along with your answers to the questions below.

✦ Include these Fathom printouts. Label each with a title and your name(s):

- Population relative frequency histogram
- Population summary table (including mean and population standard deviation)
- Relative frequency histogram for sampling distribution of the sample mean for n = 5
- Summary table for sampling distribution of the sample mean for n = 5 (including mean and population standard deviation)
- Split density histogram for n = 5 and n = 50 with superimposed normal curves
- Split normal quantile plot for n = 5 and n = 50
- Split summary table for sampling distributions of the sample mean for n = 5 and n = 50 (including mean and population standard deviation)

✦ Answer these questions:

1. What does one case in the Penny_Population collection represent? One case in the Sample of Penny_Population collection? One case in the Measures from Sample of Penny_Population collection? (It may help to select each of these collections and drag down tables so you can see the 686 cases in the Penny_Population collection, 50 cases in the Sample of Penny_Population collection, and 2000 cases in the Measures from Sample of Penny_Population collection.)

2. By looking at your population (Penny_Population) relative frequency histogram, you should recognize that the population of penny ages does not have a normal distribution. Instead, it has a geometric distribution. Practically speaking, why might the population distribution of penny ages take this form?

3. What is the mean age ($\mu$) of the pennies in your "population?" What is the standard deviation ($\sigma$)? (Answer these questions by consulting the appropriate summary table.) Round to 1 decimal place.

4. Look at the density histograms for your two sampling distributions. As the sample size increases, what is the effect on the shape of the sampling distribution? Discuss the effects of sample size on both the center and the spread of the sampling distribution.

5. According to section 8.2 in your textbook, what is the general relationship between $\mu$ and $\bar{x}$? Did your simulation results for n = 5 and n = 50 display this relationship? Explain. Be sure to cite the relevant summary table results. What is the general relationship between $\sigma$ and $\sigma_{\bar{x}}$? Did your simulation results for n = 5 and n = 50 display this relationship? Explain. Be sure to cite the relevant summary table results. Why might your results have been somewhat off? In other words, how does the theoretical definition of a sampling distribution differ from the actual simulation procedure you used? (Consult my sampling distribution lecture slides on Blackboard for help answering this question.)

6. How well does an overlaid normal curve fit each of your sampling distributions? A normal quantile plot is a more accurate method for determining how close to normal a sampling distribution is. In short, the closer the normal quantile plot is to a straight line, the better a normal distribution "fits." Judging from your normal quantile plots, which of your sampling distributions was closer to normal? Explain your reasoning. What is the general relationship between sample size and how well a sampling distribution of $\bar{x}$ can be described by a normal distribution?
7. I generated the following sampling distributions of the sample proportion ($\hat{p}$) using the same Penny_Population collection that you used for your Fathom project. First, I found the proportion of pennies in the population that were one-year old. Then, I drew 1000 samples for each of three different sample sizes (10, 100, and 200) and calculated the proportion of pennies in each sample that were one-year old. Finally, I plotted a density histogram for each sampling distribution and generated summary statistics. The vertical lines depict the mean ($\mu_{\hat{p}}$) of each sampling distribution. Consult my Fathom output to answer the following questions.

<table>
<thead>
<tr>
<th>Penny_Population</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 = proportion (Age = 1)</td>
<td>0.122449</td>
</tr>
</tbody>
</table>

Measures from Sample of Penny_Population (Proportions)

<table>
<thead>
<tr>
<th>SampleSize</th>
<th>10</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion_oneYr</td>
<td>0.1216</td>
<td>0.12132</td>
<td>0.12365</td>
</tr>
<tr>
<td>S1 = mean ( )</td>
<td>0.103216</td>
<td>0.031104</td>
<td>0.0230126</td>
</tr>
<tr>
<td>S2 = popStdDev ( )</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>S3 = count ( )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) According to the Central Limit Theorem, the standard deviation for a sampling distribution of sample proportions is $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$. Use this formula to compute the theoretical values of $\sigma_{\hat{p}}$ for each sampling distribution, and then compare these theoretical values to the simulation results displayed above in my split summary table. How well do the simulation results match the theoretical values? How might you explain any discrepancies?

b) I have labeled the sampling distribution of the sample proportion for $n = 10$. Which of the remaining two sampling distributions corresponds to $n = 100$ and which to $n = 200$? Explain how you determined which sampling distribution was which.

c) Explain why the sampling distribution of the sample proportion for $n = 10$ cannot be approximated by a normal curve.