Introduction to Sampling Distributions—Random Rectangles

You will need
- **Sampling Rectangles.ftm**

In this activity you'll use Fathom to draw samples of size 5 from a collection of rectangles many times to explore the sampling distribution of various summary statistics.

EXAMINE DATA

1. Open **SamplingRectangles.ftm.** The collection named Rectangles contains 100 random rectangles.

Q1 Make a plot of *Area* and describe the distribution of *Area* in terms of shape, center, and spread.

INVESTIGATE

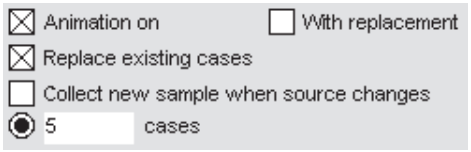
Taking a Sample


Sample of Rectangles

2. With the collection selected, choose **Collection | Sample Cases.** By default, Fathom takes a sample of ten cases with replacement and places them in a new collection named Sample of Rectangles. You'll change this to five cases without replacement.

Notice that animation is on by default. You may want to change this later.

3. Double-click the sample collection to show its inspector. On the **Sample** panel, change the settings to match these. Click **Sample More Cases.**

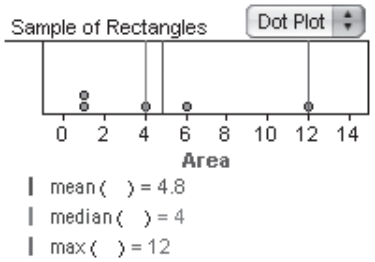


Remember that you can drag an attribute from the **Cases** panel of the inspector.

4. You now have a sample of 5 areas from the collection of all 100 rectangles. Make a dot plot of the attribute *Area* for your sample.

Choose **Graph | Plot Value** and type `mean()`. Then do the same for the other values.

5. You need to compute some summary statistics for your sample. Plot the values for the mean, median, and maximum of your sample.



Q2 In a moment, you are going to repeat this sampling, say, 200 times, to create sampling distributions of these summary statistics (or measures). But first, sketch distributions to predict what you will get for the set of 200 mean, median, and maximum areas.

Collecting Measures

Now you need to define the measures and collect them for several samples.

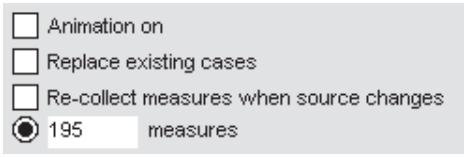
6. Double-click the sample collection to show its inspector. On the **Measures** panel, define the three measures shown. The values in the inspector should be the same as those plotted on your graph.

Measure	Value	Formula
MeanArea	4.8	mean(Area)
MedianArea	4	median(Area)
MaxArea	12	max(Area)

You can see the samples being taken by watching the dot plot.

7. Select the Sample of Rectangles collection and choose **Collection | Collect Measures.** You should see Fathom take five samples from the Rectangles collection. Each time, Fathom places the measures in a new collection named Measures from Sample of Rectangles.
8. Double-click the Measures from Sample of Rectangles collection to show its inspector. Go to the **Cases** panel. Confirm that the attributes are the measures you defined in step 6.
9. Make three histograms, one for each of the attributes in Measures from Sample of Rectangles.
10. Five samples don't make a very good distribution. Show the inspector for the Measures from Sample of Rectangles collection. Go to the **Collect Measures** panel and change the settings to match these.

Notice that Replace existing cases is off by default, so you need only 195 more measures to make a total of 200 measures.



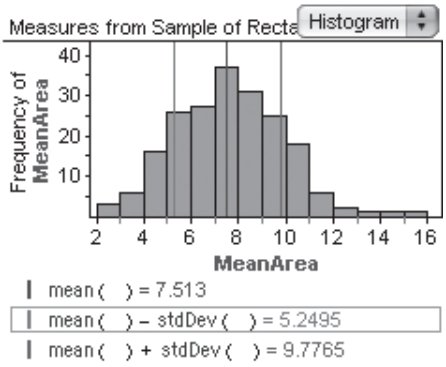
11. Click **Collect More Measures.** Even with the animation off, it will take some time to collect another 195 samples of size 5. You can see Fathom taking these samples from the Rectangles collection and calculating the measures by watching the histograms you created in step 9.

You can also select the measures collection and choose **Collection | Collect More Measures.**

Q3 Sketch your histogram of the sampling distribution for each measure.

Choose **Graph | Plot Value.**

12. Plot the mean, mean *plus* stdDev, and mean *minus* stdDev on each graph.



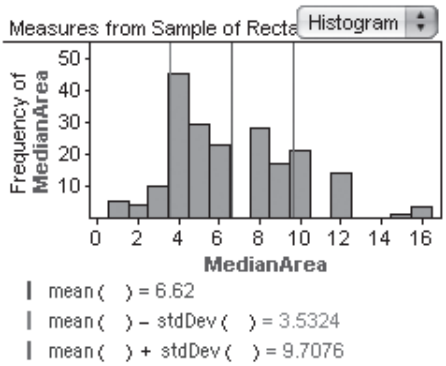
The standard deviation of the sampling distribution is often called the *standard error.*

Q4 For the sampling distribution of *MeanArea,* describe the shape, mean, and standard error. How does your sampling distribution of *MeanArea* compare to your prediction in Q2?

Q5 How does your sampling distribution of *MeanArea* compare with the population's distribution of *Area* that you plotted in Q1 in terms of shape, mean, and standard deviation?

Q6 Is the sample mean a good estimate of the mean of the areas of all 100 rectangles?

Q7 Approximately what values of the sample mean for samples of size 5 would be reasonably likely? Which would be rare events?



Q8 Repeat Q4–Q7 for the sample median. For Q6, you will need to compute the median of the population of 100 rectangle areas.

Q9 Repeat Q4–Q7 for the sample maximum. For Q6, you will need to compute the maximum of the population of 100 rectangle areas.

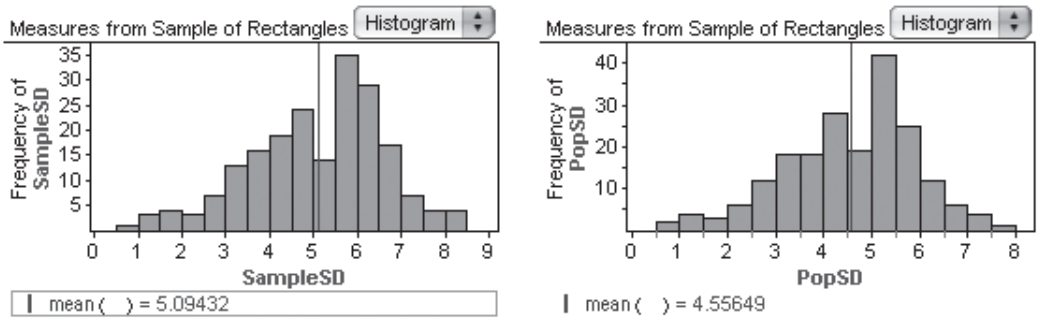
Q10 How does the sampling distribution of *MedianArea* compare to the sampling distribution of *MeanArea?*

EXPLORE MORE

1. Define two new measures in your Sample of Rectangles collection:

MaxArea	16	max(Area)
SampleSD	5.49545	stdDev(Area)
PopSD	4.91528	popStdDev(Area)
<new>		

Collect 200 measures and make a histogram for both standard deviations. Plot the center on both graphs. Compare the center to the population standard deviation, which is 5.20. Which gives a better estimate for the population standard deviation?



You'll want to turn animation off for both the sample and measures collections.

2. What will happen to the sampling distributions if you take 1000 samples rather than 200 (in the Measures from Sample of Rectangles collection)? Make your predictions and then try it.
3. What is the effect of choosing samples of size 10 instead of 5? Predict, do it, and write up the results.
4. In addition to the mean, median, and maximum, there are a lot of other summary statistics you could use. Pick one, then predict and generate its sampling distribution. Describe your results.

Introduction to Sampling Distributions—Random Rectangles

Objectives

- Understanding the concept of a (simulated) sampling distribution—the distribution of summary statistics you get from taking repeated random samples
- Identifying the characteristics of sampling distributions: The sampling distribution of the sample mean is mound-shaped and approximately normal, and the mean is at the population mean, whereas the sampling distribution of the sample median is more spread out and less mound-shaped, and the median is near, but not always at, the population median.

Activity Time: 40–50 minutes

Setting: Paired/Individual Activity or Whole-Class Presentation (use **SamplingRectangles.ftm** for either)

Statistics Prerequisites

- Familiarity with taking a sample
- Comparing distributions graphically
- Measures of center and spread

Statistics Skills

- Sampling distributions of summary statistics: mean, median, and max (SD optional)
- Definition of a sampling distribution
- The mean and SE of a sampling distribution
- Biased vs. unbiased statistics
- Identifying characteristics of the sampling distribution in terms of shape, center, and spread
- Preview of the Central Limit Theorem
- The necessity of repeated sampling

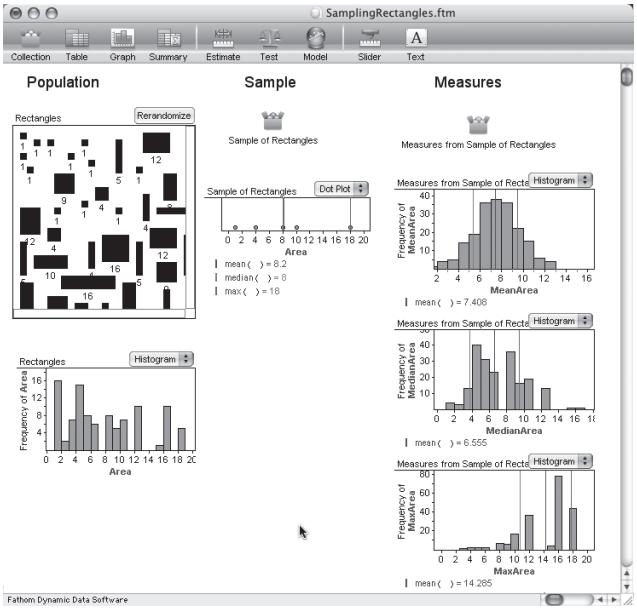
AP Course Topic Outline: Part I; Part II B (4); Part III D (1, 2, 6)

Fathom Prerequisites: Students should be able to make graphs, plot values, and define attributes.

Fathom Skills: Students sample from a collection, define and collect measures, and use the population standard deviation function.

General Notes: This activity is essential to ensure that students understand the concept of a sampling distribution. Fathom allows students to easily sample repeatedly.

Procedure: Because this is the first activity in which students generate a sampling distribution, confusion is inevitable. There are three levels of abstraction in this simulation, each level represented by a Fathom collection: the population of rectangles, a sample of five rectangles, and a collection of measures (summary statistics) that result from repeated sampling. Keeping these three levels straight is not easy. If you can project the computer screen so that everyone can see it, you can use Fathom to help students understand these levels. Divide the Fathom window into three vertical areas, one for each level, as shown here. With animation on, collecting summary statistics will cause balls to move from the population to the sample, then from the sample to the measures collection. If you use this activity as a whole-class presentation, open the Sample of Rectangles collection so that students can see which rectangles have been randomly selected for each sample.



Before you begin the activity proper, you may want to have students open **SamplingRectangles.ftm** and explore the Rectangles collection. Looking in the case table, you'll see that "behind" the display are five attributes: *Height, Width, Area, Px,* and *Py. Height, Width, Px,* and *Py* are primarily used to create the illustrated display. *Area* contains the important data, and it is the attribute that students will use for collecting measures.

EXAMINE DATA

Q1 The population has mean 7.41 and SD 5.2. The shape is not normal and is closer to uniform although it has two values (1 and 4) that rise higher than the general pattern.

INVESTIGATE

Simulated answers will vary.

Q2–Q3 The three histograms in the right column on the preceding page should be similar to what the students draw here.

Q4–Q7 The simulated sampling distribution of *MeanArea* is approximately normal with mean 7.408 and SD about 2.08. The shape of the sampling distribution is completely different from the shape of the original population, but the mean is very close to the population mean. The SD of the sampling distribution is quite a bit smaller (as it should be—theoretically it is $5.2/\sqrt{5} = 2.33$). The sample mean is a very good estimate (unbiased estimator) of the population mean because it tends to give values very close to the mean, on average. (If you took all samples of size 5, you would get that the mean of the sampling distribution was exactly the population mean.) Because the sampling distribution is approximately normal, we can estimate that reasonably likely outcomes are those in the interval $7.408 \pm 2(2.08)$, or 3.428 to 11.568.

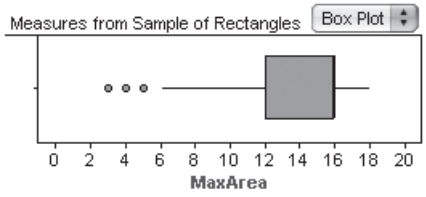
We can also estimate these values using the histogram. Rare events are those in the upper 2.5% of the distribution and the lower 2.5% of the distribution. Because there are 200 samples, this would be the largest five means and the smallest five means. We will have to approximate as we cannot isolate the five largest and five smallest from the histogram: about 12 or larger, or 3 or smaller. So, a mean larger than 12 or smaller than 3 from a sample of size 5 would be a rare event.

Q8–Q10 The typical approximate sampling distribution of the median has mean 6.62 and standard error 3.16. The mean is a little less than that of the population, and the spread is a lot less. It is more mound-shaped than the population, but we still would not call it approximately normal. The center of the distribution of medians should be below the center of the distribution of means. The distribution of sample medians is more spread out and less mound-shaped than the distribution of sample means.

The median of the population of 100 rectangle areas is 6, so it does appear that the median of the sampling distribution of medians, for random samples, is close to the population median. Here, in this sample, the median of the sampling distribution is 6. So, it appears to be a good estimator of the population median.

The sampling distribution is not very mound-shaped, so it is better to use the histogram to approximate the upper 2.5% and lower 2.5% of the 200 sample medians. Rare events would be the smallest five sample medians and the largest five. The first bar contains four medians and the second bar contains three, so we cannot isolate the smallest five. Sample medians less than 2 or more than 13 would be rare events. Medians between 2 and 13 are reasonably likely.

Q9 A typical simulated sampling distribution of the sample maximum for random samples of five rectangles is strongly skewed left. The maximum in the population is 18. The sample maximum underestimates the population maximum. A small sample is not very likely to contain the maximum value from the population, so the best estimate of the population maximum should be a little larger than the sample maximum. The sample maximum is a biased estimator of the population maximum and is biased in the direction of tending to be too small. The distribution is not close to being mound-shaped, so we will use the histogram to approximate the upper 2.5% and lower 2.5% of the 200 sample maximums. Sample maximums less than 6 would be rare events. Maximums between 6 and 18 are reasonably likely. Here is a box plot of the sample maximums:



DISCUSSION QUESTIONS

- What does one case in the Rectangles collection represent? One case in Sample of Rectangles? One case in Measures from Sample of Rectangles?
- Why is the distribution of sample medians so much bumpier than the distribution of sample means?
- Why is the distribution of sample maximums skewed left?

EXPLORE MORE

1. The effect of dividing by 4 rather than by 5 makes the standard deviation larger. When you divide by $n - 1$, the center of the sampling distribution is much closer to the population standard deviation than when you divide by n. Note that, even dividing by $n - 1$, the sample standard deviation is a biased estimator of the population standard deviation—it tends to be a bit too small. The normal distribution is not a good model for sampling distributions of standard deviations unless the sample size is very large. For smaller samples, the distribution is skewed right.

2.–3. The values should remain very close to where they were for 200 samples. The sample median area will likely get closer to the mean, and the sample maximum will likely get slightly closer to the population maximum.