The Law of Large Numbers

You will need
- a bucket of beads or a random digit table
- **BucketOfBeads.ftm**
- **BucketOfBeads 2.ftm**

The first commercial Internet providers appeared in the former Yugoslavia in 1996. A study in 2002 found that out of the total population of Internet users, about 40% started using the Internet in the last year, 30% started one to two years ago, 20% started three to four years ago, and 10% started more than four years ago. (Source: http://soemz.euv-frankfurt-o.de/media-see/newmedia/main/articles/l_bacevic.htm)

Suppose you wanted to take a sample from this population to check these proportions. How would you design the sample? How big of a sample would you need to take to get the proportions you expect? Would the sample sizes need to be different for the different percentages? In this activity you'll address these questions.

COLLECT DATA

1. Describe how you would use a random digit table or a bucket of beads to simulate selecting one person from this population of Internet users in Yugoslavia to see if they started using the Internet in the last year.

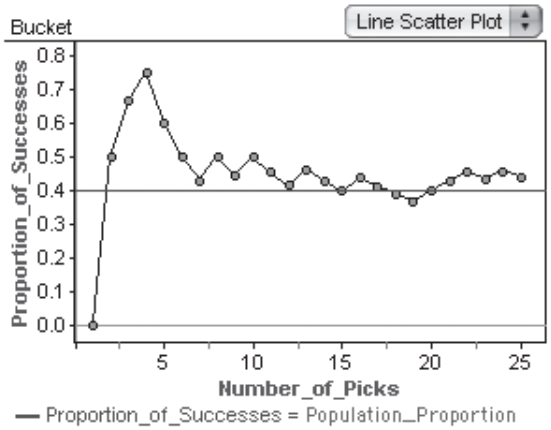
Your sample proportion will be either 0/1 or 1/1.

2. For your first pick, use your method in step 1 to select one person. Record the number of selections you have made so far (1) and whether or not your selection started using the Internet in the last year. Compute your sample proportion for the number of people in your sample who started using the Internet in the last year.

Consider a "success" to be picking someone who started using the Internet in the last year.

3. Continue picking until you have a total of 25 picks. For each pick, record the number of picks you have made so far, the *cumulative* number of successes, and the cumulative proportion of successes.

4. Plot the cumulative proportion of successes versus the number of picks.



Q1 Compare your plot with those of others in your class. Does it look like the sample proportion of successes will converge to the population proportion of successes (40%)?

GENERATE DATA

Now you'll use a Fathom simulation to investigate further whether the sample proportion of successes will converge to the population proportion of successes. First you need to build the simulation. It will be exactly like the by-hand simulation.

5. Open the Fathom document **BucketOfBeads.ftm.** You'll see a collection of 10 beads where each color represents one of the groups from Yugoslavia. There is also a slider set to the population of interest, which in this case is blue.

6. Select the Bucket collection and 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 Bucket. Drag the lower-right corner of the sample collection to see your sample.

7. Double-click the sample collection to show its inspector. On the **Sample** panel, change the number of cases to 1 and uncheck Replace existing cases.

Color_of_Interest is under Global Values in the functions list of the formula editor.

8. Go to the **Measures** panel and define three measures as shown.

Proportion_of_Successes	0.1	proportion(Ball = Color_of_Interest)
Number_of_Successes	1	count(ball = Color_of_Interest)
Number_of_Picks	10	count( )

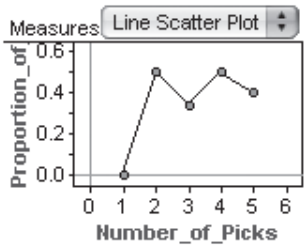
Select the sample collection. Choose **Edit | Select All Cases,** then **Edit | Delete Cases.**

9. Delete all the beads in the sample collection.

10. With the sample collection selected, choose **Collection | Collect Measures.** You should see Fathom take five samples from the Bucket collection and place a measure in a new collection named Measures from Sample of Bucket.

11. Show the inspector for the new collection. On the **Collect Measures** panel, change the number of measures collected to 1.

12. Make a plot of *Proportion_of_Successes* versus *Number_of_Picks* from the measures collection. Change the graph to a line scatter plot. Also, make a case table for the measures collection.



INVESTIGATE

Taking a Sample

Now you have everything set up to take a sample one selection at a time.

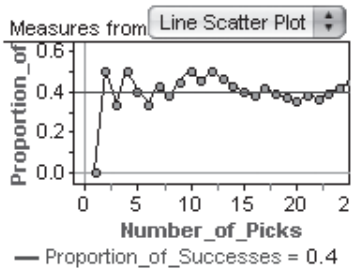
You may need to resize the sample collection to see the newest bead.

13. Click **Collect More Measures.** One more bead is picked and the measure is stored in the collection. Click **Collect More Measures** one more time.

Q2 Explain how this process of sampling in Fathom is identical to what you did in steps 1 and 2.

14. Click **Collect More Measures** until you have a total of 25 measures.

15. Choose **Graph | Plot Function** and type 0.4.



Q3 After 25 picks, what was your final proportion of successes? How close is that to 0.4?

Q4 What was the range of values you got for your proportion of successes between 1 and 25 picks?

The law of large numbers says that the difference between a sample proportion and a population proportion must be small (except in rare instances) when the sample size is large. So, you need more measures.

16. Go to the **Collect Measures** panel in the measure's inspector and change the number of measures collected to 25. Click **Collect More Measures.** Now you have 50 measures in the measures collection.

Q5 After 50 picks, what was your final proportion of successes? How close is that to 0.4?

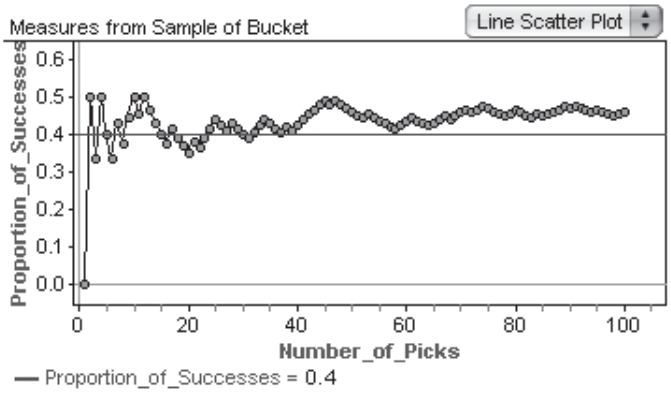
Q6 What was the range of values you got for your proportion of successes between 25 and 50 picks?

17. Repeat step 16 for 75 picks and for 100 picks. Record your final proportion of successes each time.

Q7 What was the range of values you got for your proportion of successes between 50 and 75 picks? Between 75 and 100 picks?

Use your answers to Q4, Q5, and Q7 to answer this question.

Q8 Does it look like the sample proportion of successes will converge to the population proportion of successes of 40%? Explain.



Let's see another run of 100 picks.

18. Empty the sample and measures collections.

You might want to turn Animation off.

19. Change the number of measures collected to 100. Click **Collect More Measures.** Observe the fluctuations and record the final sample proportion.

Q9 Repeat steps 18 and 19 a few times. What was the smallest sample proportion you got? The largest? Did your series of 100 picks always end up around 0.4, or did any of them surprise you?

On the slider, click on Blue and type Yellow.

20. Change the *Color_of_Interest* to yellow. Double-click the equation below the graph and change it to the appropriate proportion for yellow.

Don't forget that each time, you'll need to empty the sample and measures collections.

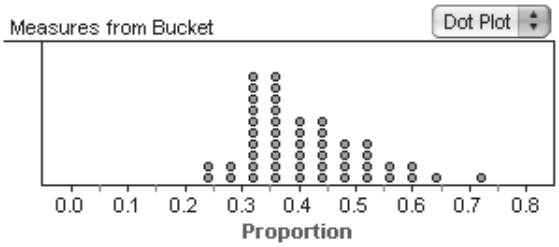
21. Run the simulation for yellow a few times. Observe the fluctuations and record the final sample proportion.

Q10 What was the smallest sample proportion you got? The largest? Did your series of 100 picks always end up around the expected population proportion, or did any of them surprise you?

Q11 Compare the simulations for blue with the simulations for yellow. How were they different? How were they the similar? Did one fluctuate more than the other?

Changing the Sample Size

Now we'd like to see bigger samples.



22. Open the Fathom document **BucketOfBeads2.ftm.** In this document you'll see a sample of 25 beads and the corresponding line scatter plot. The measures collection in this document collects 50 samples of size 25, and for each sample it records the final sample proportion and plots that proportion on a dot plot.

23. Click **Collect More Measures** a few times to see how this simulation works.

Q12 What is the range of proportions you get with samples of size 25? Describe the distribution of the sample proportion for samples of size 25 when the *Color_of_Interest* is blue.

Q13 Are the values you got in Q1 and Q3 within the range of proportions you got in Q12?

To add new cases, choose **Collection | New Cases.**

24. Add 75 cases to the Bucket collection so that the sample size is 100. Collect more measures.

Q14 What is the range of proportions you get with samples of size 100? Describe the distribution of the sample proportion for samples of size 100 when the *Color_of_Interest* is blue.

Q15 Are the values you got in Q10 within the range of proportions you got in Q13?

25. Add 900 cases to the Bucket collection for a sample of size 1000. Collect more measures.

Q16 What is the range of proportions you get with samples of size 1000? Describe the distribution of the sample proportion for samples of size 1000 when the *Color_of_Interest* is blue. Describe how the distribution of sample proportions changes as the sample size gets larger.

Q17 Does it look like the sample proportion of successes will converge to the population proportion of successes of 40%? Explain.

The Law of Large Numbers

Activity Notes

Objectives

- Designing a simulation to take a sample from a population to check given proportions
- Understanding that because of random behavior, in small samples the range of likely sample proportions will fluctuate more than in larger samples and that even in large samples there is fluctuation
- Seeing that the difference between a sample proportion and a population proportion must be small (except in rare instances) when the sample size is large or that the sample proportion of successes will converge to the population proportion of successes

Activity Time: 40–80 minutes for activity or 30–40 minutes for presentation

Setting: Paired/Individual Activity (use **BucketOf Beads.ftm** and **BucketOfBeads2.ftm**) or Whole-Class Presentation (use **BucketOfBeadsPresent.ftm**). See the Procedure section for details.

Optional Documents: Cards.ftm, Cards2.ftm (Whole-Class Presentations or Extensions)

Materials

- One random digit table per student or a bucket of beads with 4 colors (40% blue, 30% red, 20% green, and 10% yellow)

Statistics Prerequisites

- Familiarity with sampling distributions
- Some familiarity with probability distributions
- Definition of probability
- Familiarity with designing simulations

Statistics Skills

- Probability simulation
- Working with the definition of probability and probability distributions
- Counting successes versus failures
- Comparing actual data to a hypothesized model
- The law of large numbers

AP Course Topic Outline: Part III A (1, 2)

Fathom Prerequisites: Students should be able to make graphs, define attributes and measures, and work with different collections: the original, sample, and measures collections.

Fathom Skills: Students use sample and measures collections to compute cumulative proportions and values, use Fathom to create a simulation of a probability distribution, collect measures to compare models, collect measures from various size samples, delete cases, and use a slider for a non-numerical variable.

General Notes: This activity demonstrates the variability in sampling due to randomness in small and large samples. Students see that if they want to estimate a proportion, it is better to take a larger sample than a smaller one. The activity uses Fathom to repeatedly sample from a population in which outcomes are not equally likely. Students compare the effect of sample size on the sampling distribution of the sample proportion, and they also compare how these distributions change depending on the population proportion.

The data used in this activity come from "The Development of Internet in Yugoslavia" by Ljiljana J. Bacevic, (in *New media in Southeast Europe,* ed. O. Spassov and Ch. Todorov (Sofia: Southeast European Media Centre, 2003)).

Procedure: There are many ways to organize this activity. You can have your students do the whole activity or just parts of it.

For the hands-on activity, students are asked to design a simulation for sampling from the population of Internet users in Yugoslavia. They then use their design to simulate taking a sample of size 25, at each stage calculating the cumulative proportion of successes. You can have each student do his or her own simulation, or you can do the simulation as a class with one plot on the board or overhead. As a class, the bucket of beads method works best. Have a bucket with 10 beads: 4 blue, 3 red, 2 green, and 1 yellow. Sample with replacement, mixing the beads before the next student picks. As a class, keep track of the cumulative number of picks, the cumulative number of successes, and the cumulative proportion of successes. You can have each student plot his or her proportion while the next person picks, or you could wait until all selections have been made and then make the plot.

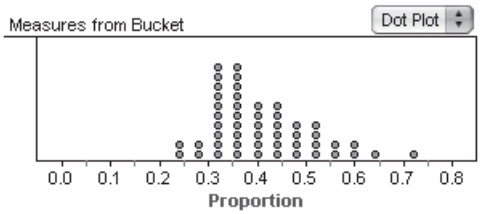
If you don't have time to use the bucket of beads idea, use the random digit table method, which goes fairly quickly.

Depending on your time, you can either have the students start with the Generate Data section and build the simulation themselves or have them start at the Investigate section (step 13) with the simulation already built. If you choose the latter, have the students start their activity with the Fathom document **BucketOfBeadsPresent.ftm** instead of the document **BucketOfBeads.ftm.** Then they can proceed with the rest of the activity as written.

The last option is that you can use the document **BucketOfBeadsPresent.ftm** or **Cards.ftm** as a presentation for the whole class. Proceed through the steps as described in the activity with either file. (Note that in **Cards.ftm** the population proportion is 0.25.)

COLLECT DATA AND INVESTIGATE

Q1–Q4 Some plots will look like the sample proportion and will converge to 0.4, but other plots might fluctuate wildly. Reasonably likely cumulative sample proportions range from 0.2 to 0.6, but as shown below, others do pop up (0.72). The range of values between 1 and 25 picks will vary greatly, depending on students' first picks. For example, for the plot on the student worksheet, the range is from 0.333 to 1; other plots could range from 0 to 0.7.



Q2 Fathom selects one bead at a time without emptying the collection first, so each measure collected is a cumulative measure. *Proportion_of_Successes* is the cumulative proportion of blue beads picked so far, *Number_of_Successes* is the cumulative number of blue beads picked so far, and *Number_of_Picks* is the total number of beads that have been picked so far.

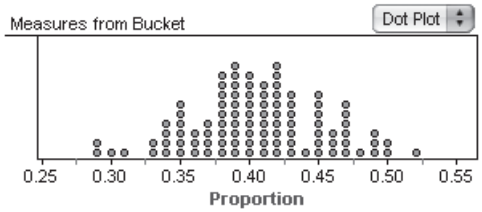
Q5–Q6 The reasonably likely range of values is 0.26 to 0.54, although smaller and larger values are possible. The range of possibilities from 25 to 50 picks will be smaller than in Q4. Typically the range will be somewhere within the band from 0.2 to 0.6.

Q7 The reasonably likely range of values for samples of size 75 is 0.28 to 0.52, although smaller and larger values are possible. The range of possibilities from 50 to 75 picks will be smaller than in Q6. Typically the range will be somewhere within the band from 0.25 to 0.55.

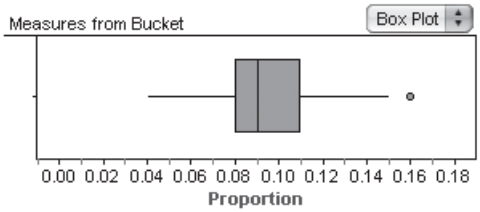
For samples of size 100, the reasonably likely range of values is 0.3 to 0.5, although smaller and larger values are possible. The range of possibilities from 75 to 100 picks will be somewhere within the band from 0.3 to 0.5.

Q8 As the sample size grows, between successive 25 marks, the range gets smaller as if it is converging. So, yes, it looks reasonable.

Q9 The answer is the same as in Q7. Here is a dot plot of 100 sample proportions. Notice there is one dot at 0.52, a rare event.



Q10–Q11 Here, a sample of size 100 will often give surprising values. Here is the distribution of sample proportions for samples of size 100 with $p = 0.1$. The median is 0.09 and the distribution is skewed right. Possible values range from 0.03 to 0.17. Typically, both will fluctuate a great deal, but the plots for $p = 0.4$ converge a little quicker.

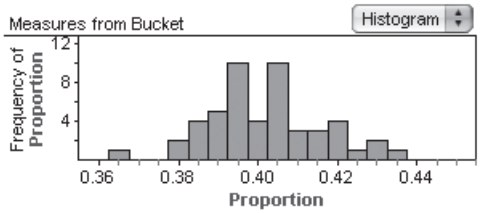


22. In this section, students generate sampling distributions of 50 sample proportions. Because 50 is a relatively small number, some of these distributions can show unexpected or rare cases, even with as large a sample size as 1000 (see plot for Q16–17).

Q12–Q13 As shown in the plot in the student activity, the distribution of the sample proportion for samples of size 25 can still be skewed. Typically, their center is around 0.4, although that too can vary.

Q14–Q15 For samples of size 100, the distribution is closer to normal, with a mean closer to 0.4 and a smaller spread (see Q7).

Q16–Q17 For samples of size 1000, the distribution should be relatively close to normal with a mean close to 0.4 and a smaller spread. Reasonably likely values range from 0.37 to 0.43. The distribution shown here has mean 0.4016 and standard deviation 0.014 and is somewhat skewed right.



EXTENSIONS

1. Have students delete the cases in the Bucket collection, then add 25 cases to the Bucket collection and change the *Color_of_Interest* to yellow. Explore how the distribution of sample proportions changes as the sample size gets larger. How is it different from the sampling distribution of sample proportions for blue?
2. Students could explore one of the other colors in the Bucket collection. Explore how the distribution of sample proportions changes as the sample size gets larger. How is it different from the sampling distribution of sample proportions for blue or for yellow?
3. Use either **Cards.ftm** or **Cards2.ftm** as a whole-class presentation. What population proportion is being investigated in these documents?