Exploration 2.1: Sampling Trees

Sampling from a Finite Population: Proportions

LEARNING GOALS

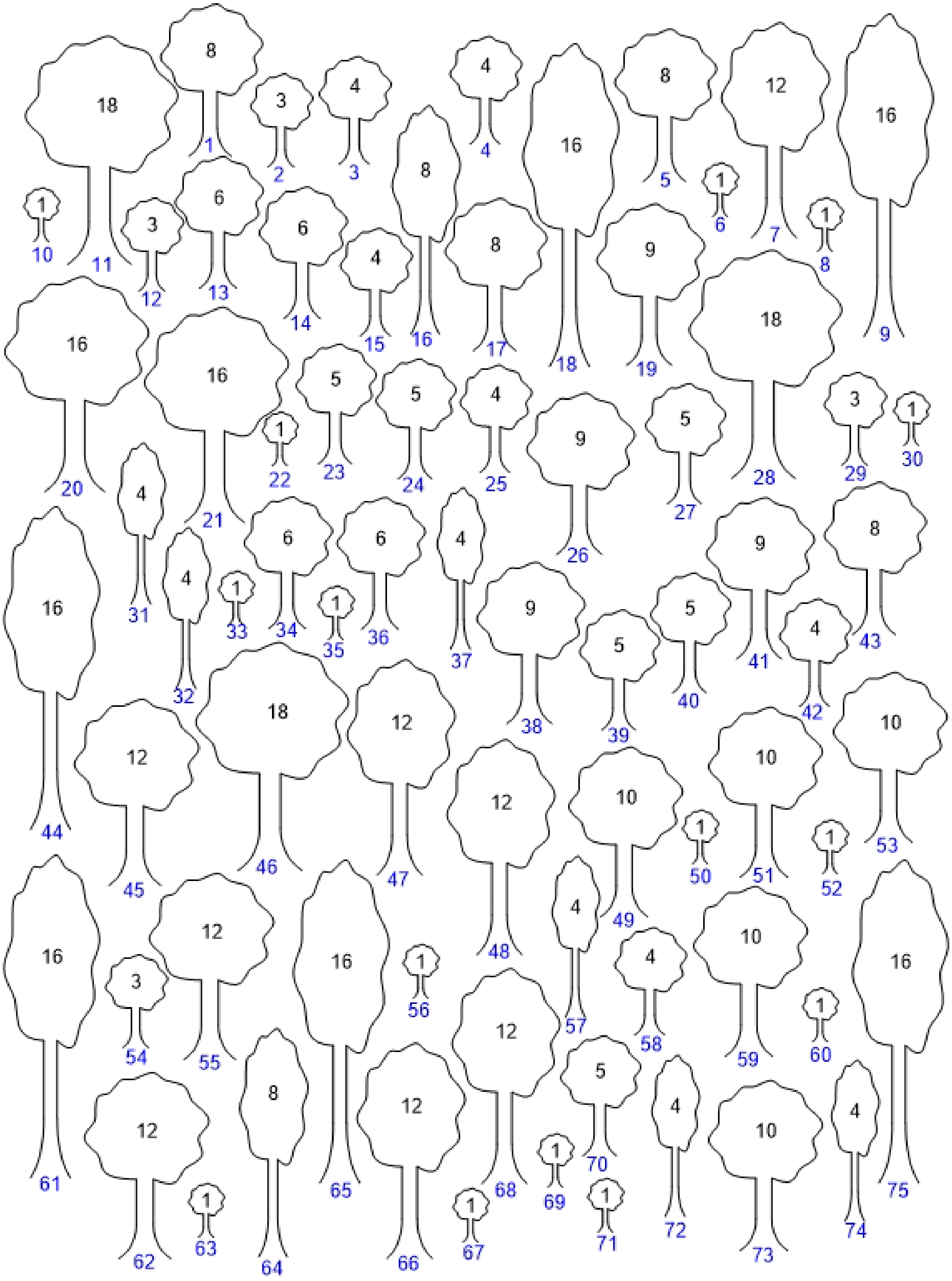
- Identify parameters (long-run proportion) and statistics (sample proportion) in a statistical study.
- Identify which statistics (proportions) and graphs (bar graph) are appropriate for categorical variables, construct graphs and calculate statistics with use of technology, and interpret appropriately.
- Identify the (finite) population and the sample in a statistical study.
- Identify whether a sampling method is likely to be biased and explain the potential impact.
- Describe how to select a random sample and recognize that one advantage of a random sample is that it is likely to be representative of the population regardless of sample size.
- Fill in a data table where rows are the observational units and columns are the variables.
- Predict the mean, standard deviation, and shape of the sampling distribution of a sample proportion from a random sample of size n, where the population proportion π is known.
- Apply simulation- and theory-based inference methods for a population proportion to research studies involving random samples from finite populations.
- Identify whether a study may be impacted by non-sampling concerns and explain the potential impact.

1. Imagine that the picture on the following page represents a grove of young maple and beech trees of various sizes. The circumference of the trunk of the tree, measured in inches, is shown in black in the canopy or crown of the tree. The tree's identification numbers are in blue at each tree's base. Select, what you think, is a representative sample of 10 trees and complete the table below by recording the circumference of each tree (the number in the canopy) and whether the tree is small (4 inches or smaller).

Tree	**Circumference** **(As noted by the black numbers in the crown or canopy of the tree)**	**Small?** **(4 in. or smaller)** **(Y or N)**
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

2. Identify the observational units and the variables you have recorded on these observational units.

These materials were developed by the STUB Network and supported by the National Science Foundation under Grant NSF-DBI 1730668. They are covered under the Creative Commons license BY-NC which allows users to distribute, adapt, and build upon the materials for noncommercial purposes only, and only so long as attribution is given to the STUB Network.


8
1
3
2
4
3
4
4
8
5
1
6
12
7
1
8
16
9
1
10
18
11
3
12
6
13
6
14
4
15
8
16
8
17
16
18
9
19
16
20
16
21
1
22
5
23
5
24
4
25
9
26
5
27
18
28
3
29
1
30
4
31
4
32
1
33
6
34
1
35
6
36
4
37
9
38
5
39
5
40
9
41
4
42
8
43
16
44
12
45
18
46
12
47
12
48
10
49
1
50
10
51
1
52
10
53
3
54
12
55
1
56
4
57
4
58
10
59
1
60
16
61
12
62
1
63
8
64
16
65
12
66
1
67
12
68
1
69
5
70
1
71
4
72
10
73
4
74
16
75

Definition

The ***population*** is the entire collection of observational units of interest, and the ***sample*** is a subset of the population on which we record data.

In this very simplistic version, we are considering the 75 trees shown as our ***population***. The 10 trees you selected are then a ***sample*** from this population. As before, we can use $\hat{p}$ to refer to a sample proportion and π to refer to the parameter, in this case the proportion of all trees in the population that are small

3. When the value of the population proportion π is unknown, we use a sample proportion to estimate its value. What is the value of $\hat{p}$ for your sample?

4. Do you think your value of $\hat{p}$ is a good estimate for the value of π? Why or why not?

Whereas any one sample may not produce a statistic that exactly equals the population parameter, we can evaluate a sampling method by seeing how the method performs across many samples.

5. Combine your results with your classmates' by producing a dotplot of the distribution of the proportion of small trees in your samples.

6. What label should be along the horizontal axis of this graph? How many dots do you have? In other words, clarify what each dot represents. (*Hint*: If you wanted to add another dot to the graph, what would you need to do?)

7. Is there sample-to-sample variation in the sample proportions in your class? What would we like to be true about this distribution to suggest the sampling method we have used here is appropriate for estimating the population proportion?

Definition

The sampling method is ***biased*** if, when using that sampling method, statistics from different samples consistently overestimate or consistently underestimate the population parameter of interest.

Note that bias is a property of a sampling *method*, not a property of an individual sample. Also note that the sampling method must *consistently* produce nonrepresentative results in the same direction in order to be considered biased.

In order to know whether a sampling method is biased, we have to know the actual value of the parameter. For the population of trees above, the proportion that are small turns out to be $\pi = 0.413$.

8. Does asking you to quickly select a representative sample of 10 trees appear to be a biased or unbiased sampling method? If biased, what is the direction of the bias (tendency to overestimate or underestimate the proportion of small trees)?

9. Explain why we might have expected this sampling method (asking you to quickly pick 10 representative trees) to be biased for this variable.

10. Do you think that if we'd asked each of you to select 20 trees instead of 10 trees it would have helped with this issue? Explain.

11. Suggest another technique for selecting 10 trees from this population in order for the sampling method to be unbiased in estimating the proportion of small trees.

Taking a simple random sample

Key Idea
A ***simple random sample*** ensures that every sample of size *n* is equally likely to be the sample selected from the population. In particular, each observational unit has the same chance of being selected as every other observational unit.

The key to obtaining a representative sample is using some type of *random* mechanism to select the observational units from the population, rather than relying on ***convenience samples*** or any type of human judgment.

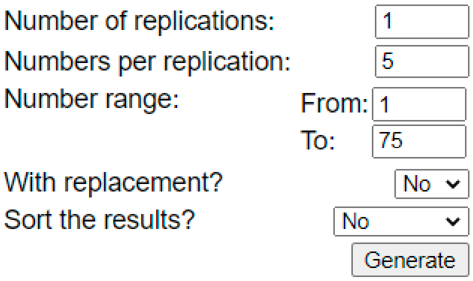
Definition
A ***convenience sample*** is a nonrandom sample of a population.

Instead of having you choose "random" trees using your own judgment, we will now ask you to take a simple random sample of trees and evaluate your results. The first step is to obtain a ***sampling frame***—a complete list of every member of the population where each member of the population can be assigned a number. Each of the 75 trees were assigned the blue numbers at their bases using labels 1-75.

12. Let's randomly select 5 trees. To do this, go to the Random Numbers applet

Specify that you want 5 **Numbers per replication** in the **Number range** from 1 to 75.

Press **Generate** to view the 5 random numbers. Enter the random numbers in the table below:



Number of replications: 1
Numbers per replication: 5
Number range: From: 1 To: 75
With replacement? No
Sort the results? No

Using your randomly generated ID numbers, look up the corresponding tree from the sampling frame above. Fill in the ***data table*** below.

Tree	**Circumference** **(As noted by the black numbers in the crown of the tree)**	**Small?** **(4 in. or smaller)** **(Y or N)**
1		
2		
3		
4		
5		

13. Let’s examine your sample of five trees, as well as those of your classmates.

    a. Calculate the proportion of small trees in your random sample.

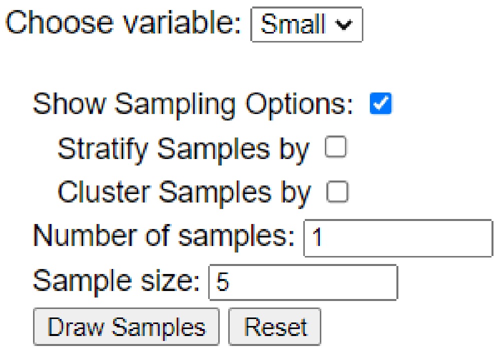
    b. Again produce a dotplot of the distribution of proportions of small trees for your sample and those of your classmates. Make sure you label the horizontal axis appropriately.

    c. Comment on how this distribution compares to your class’s dotplot of sample proportions from earlier which used nonrandom sampling.

    d. Does this second method appear to be an unbiased sampling method? How are you deciding?

Let’s use technology to take even more random samples.

- Open the **Sampling Trees** applet
- Confirm the population proportion is 0.413.
- Check the **Show Sampling Options** box. Specify 5 in the Sample Size box and press the **Draw Samples** button.



Choose variable: Small

Show Sampling Options: ☑

Stratify Samples by □

Cluster Samples by □

Number of samples: 1

Sample size: 5

You will see the ID numbers of the five selected trees appear in the box. You will also see five blue dots within the population distribution representing the five trees you have sampled (and whether they are small). These five outcomes for your sample are also displayed in a dotplot in the middle graph. The blue triangle indicates the proportion of trees in this sample that are small. This value also appears on the graph on the right.

Press **Draw Samples** again.

14. Did you get the same sample this time? Did you get the same sample proportion?

15. From the Proportion graph on the right, what is the average of the two sample proportions you have selected so far?

16. Suppose you take 1,000 different random samples. Where do you think the Proportion graph will be centered? What shape do you think it will have? Predict the largest and smallest values.

17. Change the **Number of samples** from 1 to 9998, for a total of 10,000 samples and press **Draw Samples**. Describe the behavior of the distribution of sample proportions and contrast the distribution with your predictions in the previous question, as well as to the previous class distribution of nonrandom samples using 10 trees.

Definition

A ***sampling distribution*** is the distribution of a statistic for all possible samples of size *n* randomly selected from the same population. By taking lots of samples, you can approximate this sampling distribution.

The table below outlines the simulation process you are using to approximate the sampling distribution of sample proportions of small trees from the woodlot.

Parallels between population distributions and sampling distributions

Population	All 75 trees in the woodlot
Parameter	Proportion that are small (π = 0.413)
Sample	Each sample consisted of 5 randomly selected trees
Statistic	Proportion of sampled trees that are small ($\hat{p}$ varies from sample to sample)

One big change between the random samples and the previous convenience samples is the center of the sampling distribution should now be close to π, the population proportion. In other words, simple random sampling is an unbiased sampling method because there is no tendency to over- or under-estimate the parameter.

Key Idea

The mean of the sampling distribution of sample proportions for simple random samples from a population with success proportion π will be equal to π.

As we saw, this sampling method (taking 5 randomly selected trees) is a better sampling method (unbiased) even though we selected fewer trees.

Sampling Distribution of Sample Proportions

Based on what you learned about sample proportions in Chapter 1, you shouldn't be too surprised that the distribution of sample proportions you produced in the previous question is somewhat symmetric (because π is close to 0.50), with a mean close to 0.413 (because the sampling method is unbiased). In fact, our previous formula for the standard deviation of sample proportions, $\sqrt{\pi(1-\pi)/n}$, is often still a reasonable approximation as well!

18. Verify that the standard deviation of your 10,000 sample proportions is close to $\sqrt{\pi(1-\pi)/n}$. (*Hint*: What values are you using for π and *n*?)

Key idea

When sampling from a large population, the standard deviation of sample proportions is estimated by the same formula, $\sqrt{\pi(1-\pi)/n}$, where π represents the proportion of successes in the population and *n* represents the sample size. The population is considered large enough when it is more than 20 times that sample size.

Notice one new aspect here compared to Chapter 1, when we are sampling from a finite population rather than from an infinite random process, we want the population to be large in order to assume π

represents a constant probability of success. In this case, our population is not considered large because 75 is not more than 20 × 5 = 100. For this reason, the SD of your 10,000 samples was probably a little smaller than what the formula predicts.

19. To get a larger population, we can just make more copies of the 75 trees that we have. To do this in the applet click on the **×100** radio button. This will make 100 copies of the 75 trees so now you will be sampling from a population of 7,500 trees. Take at least 10,000 samples of size 5 from this population. Is the SD of your samples now closer to what the formula predicts? Does anything else change much (shape, mean, largest value/smallest value) in the distribution of sample proportions with this larger population of 7,500 trees?

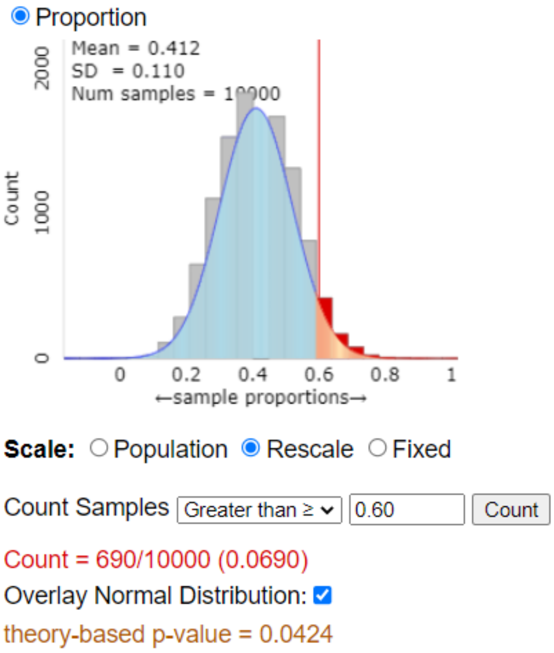
20. Assuming the population size is large enough to not impact the sample-to-sample variation, name two things that will still impact the standard deviation of the distribution of sample proportions.

21. In the **Sampling Words** applet (still with the **×100** radio button selected), select the radio button labeled **Fixed** below the distribution of sample proportions. Then change the **Sample size** to 25 and take at least 10,000 samples. Press **Draw Samples** and describe the behavior of the new sampling distribution. Does the standard deviation change as you predicted? Is the sampling method still unbiased? How are you deciding?

Notice that the shape of the distribution of sample proportions looks more like a normal distribution (still symmetric but more "filled in") when the sample size was increased, just as you saw in Chapter 1.

> **Definition**
>
> The ***Central Limit Theorem for sampling from a large finite population*** says that the distribution of sample proportions from repeated random samples will be approximately normal if there at least 10 successes and at least 10 failures in each sample.

22. In the following figure, we took 10,000 samples of size 20 from 100 copies of the tree population and then overlaid the normal distribution predicted by the Central Limit Theorem. We then determined the proportion of samples and the theoretical probability for a sample proportion of small trees of 0.60 or larger. Does the normal approximation appear to be valid here? Which p-value (0.0609, or 0.0424) do you trust more? Explain your reasoning.



● Proportion

Mean = 0.412
SD = 0.110
Num samples = 10000

Scale: ○ Population ● Rescale ○ Fixed

Count Samples Greater than ≥ 0.60 Count

Count = 690/10000 (0.0690)

Overlay Normal Distribution: ☑

theory-based p-value = 0.0424

Other Considerations

Another key property of the simple random sampling method is that we had a list of every tree in the population. This list is referred to as a ***sampling frame***. An incomplete sampling frame (e.g., only a portion of all the trees), could also produce a biased sampling method if those not in the sampling frame were systematically different from the rest of the population that is in the sampling frame.

Key Idea

If a sampling frame does not include every member of the population (e.g., trees not in the sunniest corner of the lot, trees closest to the river), then we can't claim to represent those segments in the population with our sample. An incomplete sampling frame can also lead to biased sampling.

You also need to be sure there are not any other ***nonsampling concerns*****.** These represent problems that arise even with a carefully selected sample. For example, suppose we were not able to accurately measure the circumferences of the trees, or the circumferences were measured at different heights on the tree trunks. Also, different individuals could have used different definitions of "small."

Key Idea

In addition to carefully selecting the sample, you need to guard against other possible sources of bias as well. For example, when talking to people, their suspect memories and truthfulness can impact people's answers. Other examples of ***nonsampling concerns*** include the effects that the wording of a question can have on people's responses (e.g., how extreme the phrasing is), the effects that the interviewer can have by virtue of their demeanor, sex, race, and other characteristics, using an uncalibrated scale, etc.