## Exploration 2.2: Sampling Trees (continued)

### Quantitative Data

**LEARNING GOALS**

- Describe the distribution of a quantitative variable in terms of shape, center, variability, and unusual observations.
- Use the skewness of the distribution to predict the relationship between the mean and median.
- Identify parameters (such as long-run/population means or medians) and statistics (such as sample means or sample medians) in a statistical study.
- Predict the mean, standard deviation, and shape of the sampling distribution of a sample mean from a random sample of size *n*, where the population mean, μ and population standard deviation, σ are known.

Let's reconsider the *population* of trees you first looked at in Exploration 2.1. Suppose instead of looking at the proportion of small trees (trunk circumference of 4 inches or less), we look at the data on the *circumference of the tree trunks*.

1. Is trunk circumference of the trees a quantitative or categorical variable? Why?

2. Suggest a *statistic (a single number)* you could calculate from your sample that summarizes the trunk circumference variable.

3. What *parameter* are you hoping to estimate with your statistic? (*Hint*: Be clear how this relates to the *population.*)

4. How could we explore whether the sampling methods we used in Section 2.1 are unbiased in estimating the parameter you identified in the previous question? What does it mean to be unbiased in this context?

**Definition**

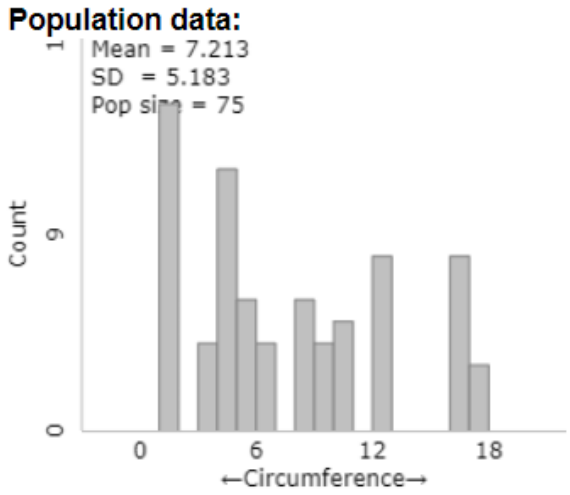
A distribution of data is ***skewed*** if it is not symmetric and, instead, the bulk of values tend to fall on one side of the distribution with a "longer tail" on the other. Right-skewed distributions have their tail on the right and left-skewed distributions have their tail on the left.

With skewed data, we might prefer to measure the "center" of the distribution with the median rather than the mean.

**Definition**

The ***median*** is the middle data value when the data are sorted in order from smallest to largest.

The *histogram* below shows the circumference of all 75 trees in the population.




5. The mean circumference in the population is 7.213 inches. Do you think the median is larger or smaller than 7.213 inches? Briefly explain your reasoning.

6. Summarize the behavior of the population shown in the histogram above. Recall that when we have a quantitative variable, we should talk about shape, center (using the mean and/or the median), variability (using the standard deviation), and outliers, as well as making sure that we are making comments relevant to the context of the study.

**Key Idea**

To distinguish a quantitative variable from a categorical variable, we will use different symbols to refer to the statistics and parameters. In particular,

| | |
|---|---|
| Statistics | $\bar{x}$ = sample mean |
| | $s$ = sample standard deviation |
| Parameters | $\mu$ = population mean |
| | $\sigma$ = population standard deviation |

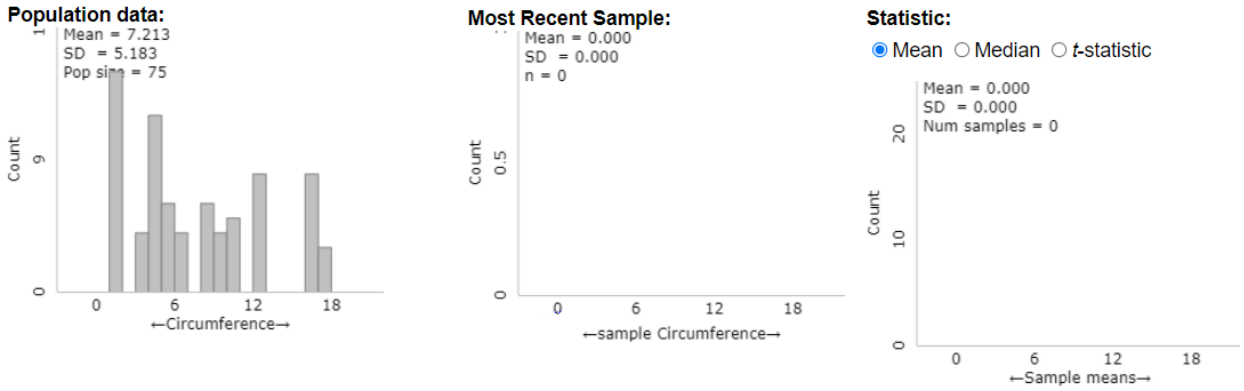
Note that we are focusing on the mean and the standard deviation. (See Section 2.4 for more discussion about the median and Chapter 6 for discussion of another measure of variability.)

7. Identify the values of $\mu$ and $\sigma$ for the tree population (from the histogram).

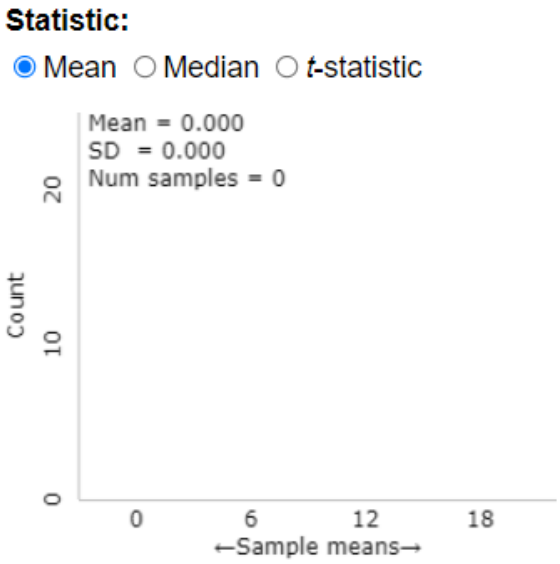
To evaluate our sampling method for a sample mean, we want to see whether the means from different samples center around the population mean.

8. Open the **Sampling Trees** applet. Use the pull-down menu to change the variable to **Circumference**. Check the **Show Sampling Options** box. Specify 5 in the Sample Size box, press the **Draw Samples** button, and choose Population for the **Scale** of the horizontal axis of the graph on the right. Notice that the graph on the left in the applet (and shown below) is the population, the graph in the middle is your sample, and the mean of this sample is placed in the

graph on the right. Fill in your results for the middle and right most graphs in the figure below. Also indicate the value of the sample mean that you obtained.




9. If we were to repeat the process in the previous question thousands of times, predict the behavior of the distribution of sample means (i.e., the sampling distribution) by describing the shape, center, and variability compared to the population. Sketch your prediction below, changing the horizontal axis scale if necessary.




10. Change the **Number of samples** in the applet from 1 to 9999 (for 10,000 total samples) and press **Draw Samples**.

    a. Sketch the distribution of sample means—including carefully labeling the horizontal axis and indicating what each dot on the graph represents. (*Hint*: How many dots do you have? What would you/the applet do to add one more dot to the graph?)

    b. Describe the shape of the distribution of sample means. How does it compare to the shape of the population?

c. What is the mean of the distribution of sample means? How does it compare to the mean of the population?

d. What is the standard deviation of the sampling distribution of sample means? How does the standard deviation of the sampling distribution compare to the standard deviation of the population?

11. Does simple random sampling appear to be an unbiased sampling method for the sample mean? How are you deciding?

12. Suppose we wanted less sample-to-sample variation in the sample means. What would you change about the sampling process?

**Key Idea**

When taking random samples from a large population (more than 20 times the size of the sample), the distribution of sample means will:

- Center around the population mean ($\mu$)
- Have a standard deviation smaller than the population, $SD(\bar{X}) = \frac{population\ standard\ deviation}{\sqrt{sample\ size}} = \frac{\sigma}{\sqrt{n}}$
- Be approximately normal when the sample size is large (by the **Central Limit Theorem**) or if the population distribution is normally distributed.

Like with proportions, we see that the statistic ($\bar{x}$) is unbiased for estimating the population mean, but now we have a different formula for predicting the standard deviation of the statistic.

13. Use your applet output to verify that the standard deviation of sample means is approximately equal to $\sigma/\sqrt{n}$. If the standard deviation is off a little bit, what is the explanation?

The Central Limit Theorem tells us that we can also assume that the distribution of sample means is approximately normal when the sample size is large. How large is large? We can't check the number of successes and failures because we are working with a quantitative variable. But in a similar manner, the sampling distribution will look normal as long as the population distribution is not too skewed. As we increase the sample size, the distribution of sample means will look more and more like a normal distribution. Generally, if the sample size is at least 20 we can assume that distribution of sample means is approximately normal. We also have a new case—the distribution of sample means will behave like a normal distribution for *any* sample size, *if* the population shape is itself approximately normal.

However, our population size wasn't more than 20 times the sample size, so the standard deviation of the sample means found in the applet might be a little lower than what is predicted.

14. If we change the sample size from 5 to 10, predict how this will impact the shape, center, and variability of the distribution of sample means.

15. Generate 10,000 random samples of 10 trees. Sketch the resulting distribution of sample means, remembering to carefully label the horizontal axis. Were your predictions about how the increased sample size would impact the shape, center, and variability accurate?

16. Using what you have seen so far, which is *least* likely:
    - Randomly picking one tree from the population and finding it to be 8 inches or more in circumference.
    - Randomly picking 5 trees from the population and finding the mean of the 5 trees to be 8 inches or more in circumference.
    - Randomly picking 10 trees from the population and finding the mean of the 10 trees to be 8 inches or more in circumference.

    Explain your reasoning.