ACTIVITY 7

STUDY HOURS

# DIRECTIONS

Have one group member upload this docx file to your Google Drive and share it with your groupmates and your instructor. Name your document “**GroupX\_Activity7”** (where X is your group number). Work together to type up your responses to each question. Download your document as a PDF and submit this to Canvas individually.

# INTRODUCTION

How many hours do you study for your classes in a typical week? Do you think you study more or less than other students? What do you think is the typical or average number of hours studied for a student at UIUC? In today’s activity, we will explore this by collecting data in our classroom. First, ask this question to yourself, and determine the number of hours you study in a typical week. Once you’ve done that, submit your data to this google form here. We’ll use this later in the class!

<https://docs.google.com/forms/d/e/1FAIpQLSf9R0thCtVp7DM9gMUuRSyhDs0i5nKcgiwMDVgbFTqHGOeuuQ/viewform>

# DATA COLLECTION

Merge your small groups together to create a group of roughly 8-12 students. In these large groups, work on creating a table of data that contains all of the numbers of hours you study. To do this, create a data table in TinkerPlots like the one below:



We’ll use this data to investigate our original research question today to determine the average number of hours a UIUC student studies!

1. Plot the sample data for your group and copy it below. Describe the distribution of your sample. Be sure to describe the center, spread/variation (use stdDev() in TinkerPlots!), and the shape of your distribution.
2. Using only the data at hand, what would you estimate the average number of study hours to be for all UIUC students?
3. Do you think your answer to the question in #2 is accurate? Why or why not? What sources of variation could be present here?

# MODELING WITH BOOTSTRAPPING

In the last activity on kissing, you learned about bootstrapping to estimate a percentage or proportion. This statistical process also works for estimating a mean or average! Review the concept of bootstrapping from our last class, and try to apply it here to create a TinkerPlots sampler that bootstraps this sample.

1. Create a TinkerPlots sampler that bootstraps your study hours data.
	* Copy and paste the sampler here.
	* Are you sampling with or without replacement? Explain your reasoning.
	* What did you set draw to and why?
	* What did you set repeat to and why?
	* Describe what happens when you run the sampler.
2. Plot the results from a single trial. Describe what a dot in your plot represents here.
3. What will you collect statistics on and why?
4. Collect on your statistic identified in the previous question 500 times and paste that plot below. Interpret what a dot in this plot represents here.
5. Where is the graph above centered, and why does that center make sense given the assumptions made when bootstrapping?
6. Compute the standard deviation (stdDev() in TinkerPlots) of your collected statistics in your sampling distribution.
	* What value did you get?
	* What does the standard deviation mean in the context of a sampling distribution here?
	* How does the standard deviation compare to the value you computed for your group’s original sample in question 1?
7. Use your divider tool:
	* Highlight the middle 95% of the sample means in your sampling distribution. What are the endpoints of this interval that covers the middle 95% of sample means? Copy and paste a picture of your plot below.
	* Highlight the middle 90% of the sample means in your sampling distribution. What are the endpoints of this interval that covers the middle 90% of sample means? Copy and paste a picture of your plot below.
	* Highlight the middle 99% of the sample means in your sampling distribution. What are the endpoints of this interval that covers the middle 99% of sample means? Copy and paste a picture of your plot below.
8. Write a sentence to interpret your 95% confidence interval in the context of the problem.

# CONFIDENCE INTERVALS, STANDARD ERROR AND THE MARGIN OF ERROR

The standard deviation of a sampling distribution has a special name called the **standard error**. This is interpreted very similarly to the standard deviation, but the only thing that changes is the observational unit, that is, what the dot in the plot represents. Standard deviation tells us how far one data value (one student’s study hours) is from the mean. Standard error works similarly, but one data value is now the sample mean. **So standard error tells us how far away our sample mean typically varies from the mean for the whole population.** Standard error can also be computed by taking the standard deviation of the sample data (rather than the collected statistics) and dividing by square root of the sample size. (Try this yourself!)

The three intervals you found in question 9 above by finding the endpoints of the middle 95%, 90% and 99% of sample means in the sampling distribution are called **confidence intervals**. Confidence intervals are used to make statistical inferences, that is, infer what the population might look like based on our sample. We know our sample mean is likely not exactly the population mean, and thus we use ***confidence intervals to provide a range of possible values.*** By picking a large percentage to cover our sampling distribution, the interval we found is ***almost sure to cover the true population mean***. 90%, 95% and 99% are commonly used levels of confidence in statistics. In a future activity, you will learn more about what the confidence level (what 90%, 95% and 99%) means. In the remainder of this activity we will focus on the general form of a confidence interval and how to compute them.

Confidence intervals are usually reported as two values: the lower and upper values of the interval (in TinkerPlots, the values at the edges of your divider tool), with the sample statistic of the original, real world sample in the center of the interval:

**(LOWER VALUE, UPPER VALUE)**

They can also be represented by this form:

**SAMPLE STATISTIC** $\pm $ **MARGIN OF ERROR**

If we use the divider tool to pick the middle 90%, 95%, or 99% of the simulated sample means, our original sample mean will be very near the center of that interval. The random nature of sampling in TinkerPlots is the only factor that makes this not quite the exact center of the interval. The interval illustrated above conveys the main idea of interval estimation – our sample statistic should be a good estimate of the parameter, but it could by off by some amount. The **margin of error** illustrates this idea of how far off we think our sample statistic is from the true parameter *at most*. With an interval that you already have found, the margin of error can be found by determining the difference between one of the endpoints of the interval and the sample statistic at the center of the interval.

# RE-EXAMINING OUR INTERVALS

1. For the intervals you created in problem 10, find the margin of error in these intervals.
2. Based on your intervals in question 10, it seems that interval estimates with higher levels of confidence are:
	* Wider
	* Narrower
	* The same width

Explain your choice above. Why does the confidence level have an impact/no impact on the interval in this way?

1. Suppose you were to combine your data with another set of 8-12 students in the class, and create confidence intervals by bootstrapping this larger set of data. Would you expect the width of your confidence interval to be:
	* Wider
	* Narrower
	* The same width

Explain your choice above. Why does the sample size have an impact/no impact on the interval in this way? (If you are unsure, try combining your data with another group and see what happens!)

1. Eventually, your instructor will show a bootstrap simulation using all of the class’ study hours data from the Google Form. What is the 95% confidence interval found for the whole class? How does it compare to the interval you found in your group? Explain what is different, and why this is the case.