# CONNECTING INTUITIVE SIMULATION-BASED INFERENCE TO TRADITIONAL METHODS

Robin Lock[1], Kari Lock Morgan[2], and Patti Frazer Lock[1]
[1]Math, CS & Statistics, St. Lawrence University, Canton, NY 13647 USA
[2]Department of Statistics, Pennsylvania State University, State College PA 16802 USA
rlock@stlawu.edu

*Simulation-based methods have become increasingly popular as a way to introduce students to the core ideas of statistical inference. Yet most advocates of this approach still recognize the need for students to also be exposed to traditional, formula-based methods based on normal and t-distributions. As we have gained experience with building basic intuitions for inference through simulations, we have also refined methods to extend those ideas to make the connections to learning traditional methods easier and more efficient. We explore how these methods help students translate the "big picture" ideas of simulation, that apply to many parameter situations, to see the common structures of traditional methods.*

## INTRODUCTION

George Cobb, in his plenary address at the first United States Conference on Teaching Statistics (USCOTS 2005) and later paper in the first issue of TISE, Cobb (2007), challenged the statistics education community to move away from introducing the core ideas of statistical inference using traditional normal and t-based procedures. Instead he advocated using randomization-based procedures that are computer-intensive but make a stronger, more accessible connection for students to the reasoning behind statistical inference. In the often-quoted lines from Cobb's TISE abstract, "Before computers statisticians had no choice. These days we have no excuse. Randomization-based inference makes a direct connection between data production and the logic of inference that deserves to be at the core of every introductory course."

Several textbooks and curriculum projects have been developed to try to implement Cobb's ideas under the general heading of simulation-based inference (SBI). Some of the first include Lock[5] (2013), Tintle, et al (2016), Tabor & Franklin (2011), the University of Minnesota's CATALST Project (Zieffler, et al, 2012), and the OpenIntro project (Dietz, et al 2014). The Simulation Based Inference Blog (https://www.causeweb.org/sbi/) is a good place to find other resources, discussions, and tips for using this approach. In what follows we discuss making the transition from these SBI methods to the more traditional formulas and standard distributions.

## ASSUMPTIONS

As the projects above indicate, there are many ways to incorporate simulation-based methods into an introductory statistics course. We start with a few assumptions to set the stage for our discussion of the transition:

1. *We start with simulation-based inference*. We won't go into the arguments for this approach here but assume that this happens relatively early in the course and is the students first exposure to the ideas of inference.
2. *We cover lots of parameter situations.* One of the advantages of SBI is that the procedures are very general and can easily be adapted and applied to different parameters (means, proportions, differences, slope, …). Students will see bell-shaped bootstrap and randomization distributions in lots of different settings before we give those common shapes a formal name.
3. *We want students (eventually) to see traditional methods.* We recognize that students may be going to go on to other courses (e.g. research methods in a discipline), use standard statistical software, or read journal articles that use procedures based on normal and t-distribution approximations. Also, while SBI's avoidance of algebraic manipulations is a plus for many groups of students, some more formula-friendly students can gain insights into how factors (such as sample size) affect the inference process from the more traditional approaches.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org 

4. *We need good software to make the SBI procedures accessible to students.* We have developed the freely available StatKey web apps (*http://lock5stat.com/statkey*) for this purpose. Another good option is the Rossman/Chance Applet Collection (*http://www.rossmanchance.com/applets/).* Some traditional statistical software packages (e.g. R, JMP, Minitab Express) have made great strides in adding accessible SBI facilities.

EXAMPLES

Before discussing transitions, we offer two examples that illustrate the SBI approach and will extend to demonstrate moving to traditional methods.

*Example #1: Online dating app use for 18-24 year olds (bootstrap confidence interval for p)*

What proportion of 18-24 year olds in the US have used on online dating app? A Pew survey (Smith, 2016) showed that 53 of 194 young adults had used on online dating service or mobile app. This gives a sample proportion of $\hat{p} = \frac{53}{194} = 0.273$. Let's find a 95% confidence interval for the proportion of all US adults in this age category who have used online dating apps.

We construct bootstrap proportions by sampling 194 people with replacement from the original sample of young adults and finding the proportion with online dating experience in each bootstrap sample. Repeating this process for 10,000 bootstrap samples using StatKey gives the distribution shown in Figure 1. We then have two ways to obtain a confidence interval from the bootstrap proportions.

*Standard error method*: For a roughly 95% interval, estimate the standard error (SE) as the standard deviation of the bootstrap proportions in Figure 1, then add and subtract two times this value from the original sample proportion.

$$\hat{p} \pm 2SE = 0.273 \pm 2 \cdot 0.032 = 0.273 \pm 0.064 = (0.209 \text{ to } 0.337)$$

*Percentile method*: Find the endpoints that give the middle 95% of the bootstrap distributions. From Figure 1 we see these go from 0.211 to 0.335.

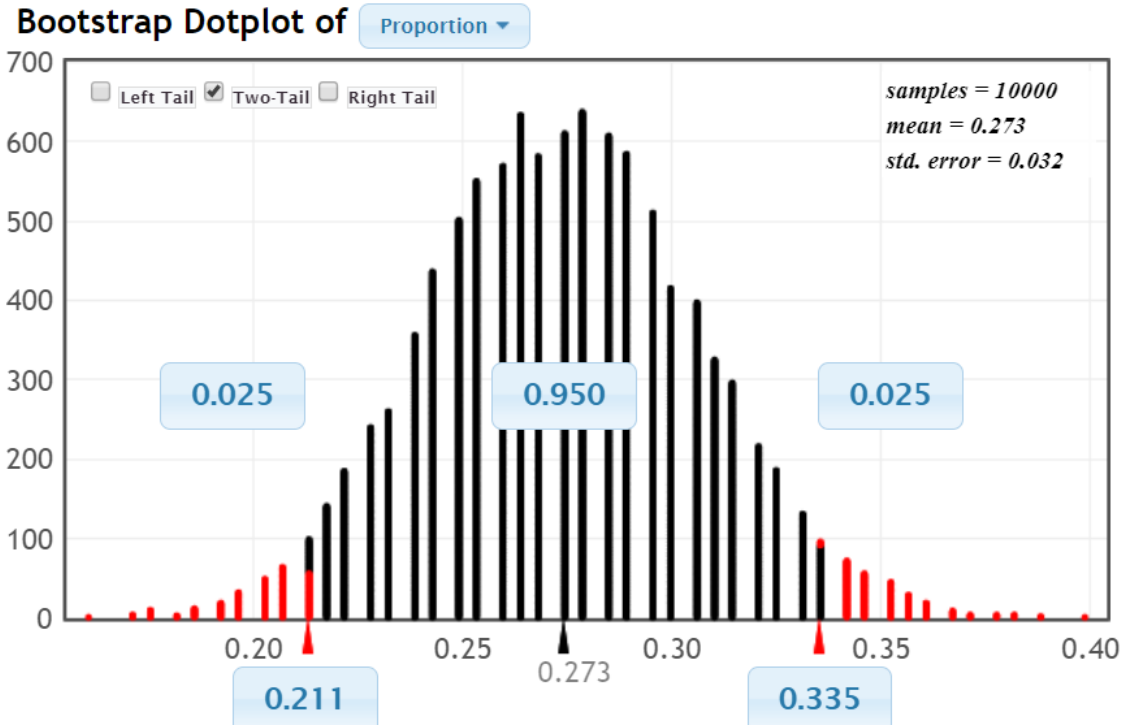


Figure 1. Bootstrap distribution of proportions of young adults using dating apps

*Example #2: Does mind-set matter? (randomization test for a difference in means)*

In an experiment (Crum & Langer, 2007) 41 female hotel maids (randomly chosen) were informed that the work they did qualified as an active lifestyle, with examples of tasks that qualified as good exercise. A different randomly chosen group of 34 maids were left uninformed about these facts. Weight loss was recorded for all participants over a four-week period. The informed group had an average weight loss of $\bar{x}_1 = 1.79$ pounds ($s_1 = 2.88$) and the uninformed group lost an average of $\bar{x}_2 = 0.20$ pounds ($s_2 = 2.32$). This gives a sample difference of $\bar{x}_1 - \bar{x}_2 = 1.59$ pounds in favor of the informed group. We want to test a null hypothesis that the population means are the same ($H_0: \mu_1 = \mu_2$) versus an alternative that the mean loss is larger for maids that are informed about the healthful benefits of their work ($H_a: \mu_1 > \mu_2$).

To do a randomization approach we scramble the 75 weight losses, re-assign them at random, 41 to the Informed group and the other 34 to the Uninformed group, then find the difference in means for the randomization sample. Repeating this process 10,000 times gives the plot shown in Figure 2. To estimate the p-value we count how many of the simulated samples had a difference in means as large as the 1.59 for the original sample. In the distribution below, that happened 47 out of the 10,000 samples to give a p-value of 0.0047, strong evidence that the mean weight loss is higher for maids who are informed of the exercise benefits of their job.

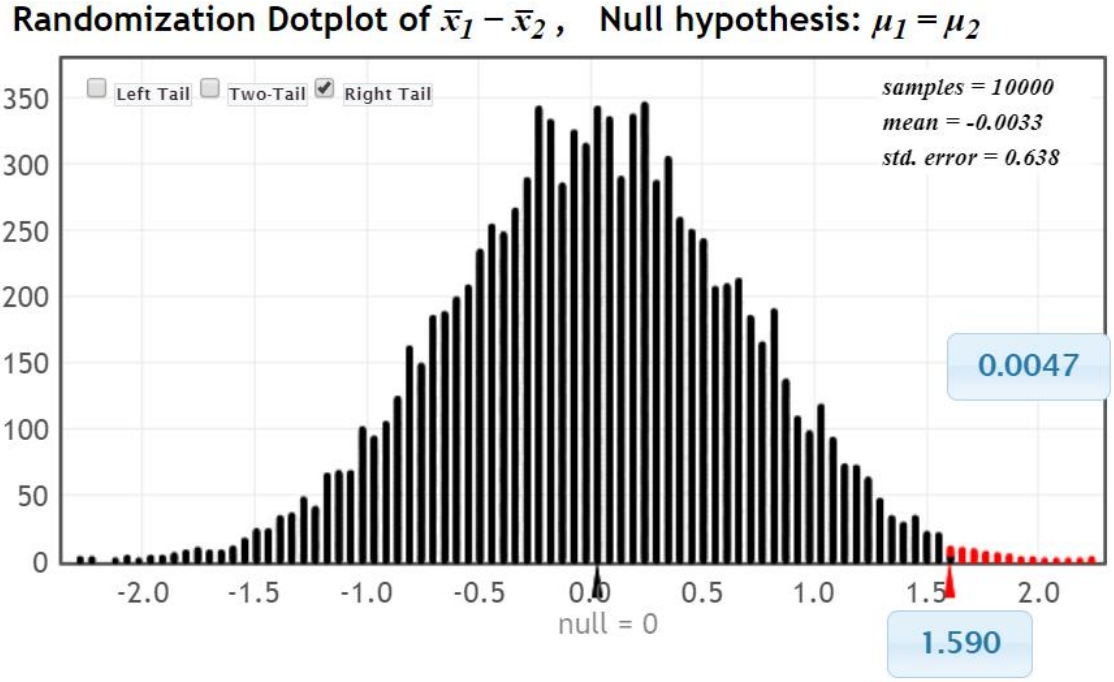


Figure 2. Randomization distribution for differences in mean weight loss

MAKING THE TRANSITION

*Step #1: Introduce the normal distribution*

Throughout the SBI material, students have been seeing lots of bell-shaped bootstrap and randomization distributions, so we introduce the idea of a normal curve to summarize that shape. They know that the bootstrap distribution should be centered at the value of the original statistic and the mean of a randomization distribution is determined by the null hypothesis. For the standard deviation of the normal, they can use the standard error as found in the bootstrap or randomization distributions. Thus for our two examples we have $\hat{p} \approx N(0.273, 0.032)$ for the proportions of young adults using online dating and $\bar{x}_1 - \bar{x}_2 \approx N(0, 0.638)$ under $H_0$ for the differences in mean weight losses. These are shown in Figure 3 along with the endpoints and areas that give the confidence interval and p-value. Note that these use the same process as the SBI methods, only substituting smooth curves for simulation dotplots and areas for counting proportions of dots in the tails. The confidence interval is very close to the bootstrap and the p-value leads to a similar strength of evidence.

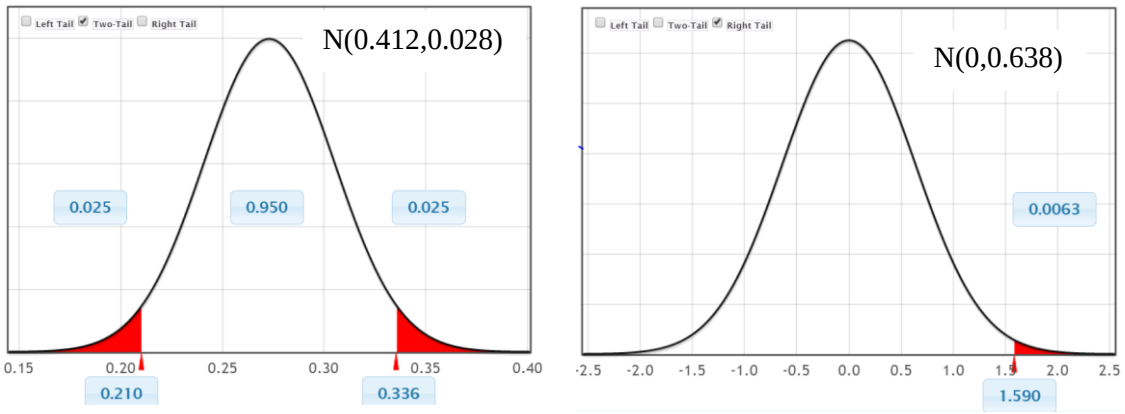


Figure 3. Normal distributions to approximate Figures 1 and 2

*Step #2: Standardize*

While it may be convenient to initially work in the scale of the original data, we often prefer to move to a standard N(0,1) scale. This eliminates the need to specify a mean and standard deviation to reset the software for each new example (not a major task) but also helps students get some intuition on what is likely to occur even before they go to software. The two key standardization formulas are

Confidence interval:

$$Statistic \pm z^* \cdot SE$$

Hypothesis test:

$$z = \frac{Statistic - Null}{SE}$$

The $z^*$ for the confidence interval comes from finding the N(0,1) endpoints that give the desired confidence level in the middle. The p-value is found as the N(0,1) area beyond the z-statistic (depending, as in the randomization, on the "tail" of the alternative hypothesis). At this stage we still use the SE as found from the bootstrap or randomization distribution (but that is about to change in Step #3).

For our online dating app example, we have $0.273 \pm 1.96 \cdot 0.028 = (0.210 \text{ to } 0.336)$, matching (of course) the result of Step #1. Now students have reinforcement for where the "2" comes from in the original standard error method and have an easy way to extend that method to other confidence levels – just find the right $z^*$ multiplier to replace the "2".

For the mind-set matters example, we have $z = \frac{1.59-0}{0.638} = 2.49$ which gives an upper tail p-value from N(0,1) equal to 0.0063 (again matching Step #1). Since all the z-statistic is doing is measuring how many SE's the sample statistic is from the null parameter, students quickly get a feel that z-values beyond 2 or 3 (in magnitude) are likely to provide significant evidence against the null while values near one or smaller are unlikely to do so.

*Step #3: Use a formula for the standard error*

So far we really haven't gained much over using just SBI techniques. If we have to construct 1,000's of bootstrap or randomization samples to estimate the SE for the normal distribution or standardization, why not just find the confidence interval or p-value directly from that simulated distribution? Of course, instructors know the answer – for many of the standard parameter situations statisticians have derived formulas which can be used to estimate the standard error from basic summary statistics without needing any simulations. That is the key to the final step.

We actually need to keep track of two questions to implement the traditional method. What is the formula for estimating the standard error? What are the conditions needed to justify using a standard distribution? Let's see how these work for our two examples.

For the confidence interval for the proportion of young adults using online dating apps, the relevant formula for the standard error of the sample proportion is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

and the distribution will be reasonably normal if the sample size is large. One easy condition for "large" is that the sample has a least ten cases with "yes" and at least ten with "no" which is clearly met for that sample. This gives the "usual" interval

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.273 \pm 1.96 \sqrt{\frac{0.273(1-0.273)}{194}} = 0.273 \pm 1.96 \cdot 0.032 = (0.210, 0.336)$$

While we don't derive the formula for SE, students can easily verify that the result is similar to what is found from a bootstrap distribution.

For the test of difference in two means with the informed/uninformed data we have

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{2.88^2}{41} + \frac{2.32^2}{34}} = 0.601$$

and we also introduce the t-distribution as an alternative to the normal when we are using standard deviations from the samples in this formula to estimate the standard error. Since these samples are relatively small, we also check that the two samples are relatively symmetric and don't have any big outliers. Computing the t-statistic we have

$$t = \frac{Statistic - Null}{SE} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE} = \frac{1.79 - 0.20}{0.601} = \frac{1.59}{0.601} = 2.65$$

Using the upper tail of a t-distribution with 33 degrees of freedom (we use the conservative "smaller of the df for the two samples" rule) we have a p-value of 0.0062. The p-value and conclusion are similar to the previous work with this example.

OBSERVATIONS

We are only using the normal distribution as a tool to approximate the familiar bell shape that students have seen in lots of bootstrap and randomization distributions. They already have plenty of experience finding endpoints and tail proportions for such curves, although in the more concrete setting of counting dots rather than finding areas. Using technology such as StatKey, that presents an intentionally similar interface for dealing with the simulation and theoretical distributions, making that part of the transition very easy for students. They have a much easier time working with the normal distribution than in pre-SBI days when normal calculations were the starting point (and even worse when such calculations were done with paper tables).

Some instructors worry that students will get confused or overloaded if we are showing them multiple ways to solve the same problem. Wouldn't it be better to just go straight to the final formula and skip the middle steps? Unfortunately, many students are not as enamored or comfortable with algebraic manipulations of formulas as their instructors. That's one of the main advantages of using SBI to introduce the ideas of inference in a general framework that does not rely so heavily on formulas and mathematical machinery. We have found that the three steps outlined above help students make the transition to formulas and, as they become comfortable with the process they can skip the first two steps. Once they have the general structures of $Statistic \pm (z^* \; or \; t^*) \cdot SE$ and $(Statistic - Null)/SE$ they can easily move to a new parameter situation by getting a new formula for the SE and knowing the proper reference distribution.

Instructors also worry about adding SBI methods to a course, while still covering traditional methods in a curriculum that is almost always too full to begin with. We have found that this is not as big an issue as one might expect. Students have already grappled with many of the important ideas of inference (e.g. how to interpret a confidence interval, how to set up

hypotheses, what a p-value measures and how to use it to interpret the results of a hypothesis test in context) during the SBI portion. When going through the traditional methods the main new ideas are the short formulas that let us estimate SE without needing thousands of simulations. We have been pleasantly surprised at how much quicker the traditional methods go when we aren't having students trying to learn the important ideas of inference at the same time they are trying to deal with these different formulas.

In most introductory courses we don't (and shouldn't) provide enough mathematical machinery to derive the formulas for various standard errors. Having experience with SBI allows students to at least verify that the results of the formula are roughly consistent with what they observe from simulations.

Consider the case of a single proportion. For a confidence interval we use $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ but for a test of $H_0: p = p_o$ we use $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$ and students wonder why the difference? After SBI the explanation is easy since the confidence interval is coming from a bootstrap distribution which students know is centered at the sample statistic, $\hat{p}$, while the test deals with a randomization distribution which is centered around the null hypotheses, $p_0$. Since the purpose of the normal curve is to approximate either the bootstrap or randomization distribution, it makes sense that the proportion used to find the SE should be the one from the center of the simulation distribution.

Although we have interspersed the examples for confidence intervals and hypothesis tests in this paper, in practice with our students, we focus on one type (hypothesis tests first) and then do the other. In earlier iterations we discussed the normal distribution (finding areas, endpoints and standardization) abstractly and then jumped to the application to confidence intervals and tests. We found that students had more trouble making the connections and that the transition goes more seamlessly if we step through using the normal distribution to approximate the simulation distribution in the respective settings.

## CONCLUSION

We agree with Cobb that simulation-based methods establish ties to the logic of statistical inference that improve understanding for students. Furthermore, using these methods as the starting point for inference paves the way for students to later more easily extend those important ideas to the more formula-driven, but still very common, traditional methods.

## REFERENCES

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum. *Technology Innovations in Statistics Education*, *1*(1). https://escholarship.org/uc/item/6hb3k0nz

Crum, A. and Langer, E. (2007) "Mind-Set Matters: Exercise and the Placebo Effect" *Psychological Science*, *18*, 165-171

Diez D.M., Barr C.D., & Cetinkaya-Rundel M. (2014). *OpenIntro: Introductory Statistics with Randomization and Simulation*. https://www.openintro.org

Lock, R., Lock, P. F., Lock Morgan, K., Lock, E., & Lock, D. (2013). *Statistics: Unlocking the Power of Data*. Hoboken, NJ: John Wiley and Sons.

Simulation Based Inference Blog https://www.causeweb.org/sbi/

Smith, A. (2016) "15% of American adults Have Used Online Dating Sites and Mobile Dating Apps", Pew Research Center, http://www.pewinternet.org/2016/02/11/15-percent-of-american-adults-have-used-online-dating-sites-or-mobile-dating-apps/

StatKey – web apps for doing simulation-based inference at http://lock5stat.com/StatKey

Tabor, J., & Franklin, C. (2011). *Statistical Reasoning in Sports*. W.H. Freeman & Company.

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2016). *Introduction to Statistical Investigations*. Hoboken, NJ: John Wiley and Sons.

Zieffler, A., & Catalysts for Change (2012). *Statistical thinking: A simulation approach to modeling uncertainty*. Minneapolis: Catalyst Press.