INTUITIVE INTRODUCTION TO THE IMPORTANT IDEAS OF INFERENCE

Robin Lock[1], Patti Frazer Lock[1], Kari Lock Morgan[2], Eric Lock[2], and Dennis Lock[3]
[1]St. Lawrence University, [2]Duke University, and [3]Iowa State University, USA
rlock@stlawu.edu

Concepts of statistical inference, such as margin of error when estimating a parameter and p-value when testing a hypothesis, are notoriously difficult for students to grasp. In traditional approaches, these ideas typically come as the culmination of a long development of prerequisite material on sampling distributions, formulas for standard errors, standard reference distributions, central limit theorems, and formulas for standardizing values. Simulation methods, such as bootstrap intervals and randomization tests, require minimal background knowledge and highlight the underlying logic of statistical inference, giving students an intuitive appreciation for the key ideas early in a course. But are such methods accessible and understandable to beginning students? We argue that advances in technology make this approach both feasible and desirable. So how does one go about modifying a course to incorporate these ideas? That question is the main focus of this paper.

BACKGROUND

George Cobb in his address to the first United States Conference On Teaching Statistics (USCOTS 2005) and subsequent lead paper in the first edition of TISE (Cobb 2007) challenged statistics educators to move away from reliance on inferential methods based on standard normal and t-distributions and embrace simulation-based randomization methods as a more authentic way to introduce the core ideas of statistical inference. Also, these methods are now part of new Common Core Standards for Mathematics in the United States. In a frequently quoted conclusion to his paper, Cobb states "*Before computers, there was no alternative. Now, there is no excuse.*"

But many statistics teachers still had some "excuses". Where could we find guidance on how to implement such an approach within an introductory curriculum? What about materials (textbooks, activities, sample assessments, etc.) to support these methods? Is there technology that is easily accessible and appropriate for use with beginning students? Should we use simulation methods exclusively or still include formula-based methods using traditional approximating distributions? How should simulation methods fit within a syllabus for an entire course?

We address some of these questions below, addressing first What, then Why, then How.

WHAT SIMULATION METHODS SHOULD WE TEACH?

We broadly categorize inference questions in an introductory statistics course into two groups: estimation and testing. These two types of inference lead naturally to two general classes of simulation procedures: *bootstrapping* to estimate sampling errors and create confidence intervals, and *randomization tests* to measure strength of evidence and find p-values for significance tests.

Bootstrap Distributions for Estimating Standard Error and Creating Confidence Intervals

Suppose we have a sample statistic estimating a population parameter. We'd like to assess how accurate that statistic is likely to be for our sample size and population. Generating multiple samples from the population is generally not feasible, so instead we generate multiple *bootstrap samples* from the original sample. To generate one bootstrap sample we choose values *with replacement* from the original sample using the same sample size. We then compute a bootstrap statistic for that sample. Repeating this process for many (1000's) of bootstrap samples gives a bootstrap distribution, which we can use to measure the variability in the sample statistic.

We show students two different ways to produce a confidence interval from a bootstrap distribution. The first uses the standard deviation of the bootstrap statistics to estimate the *standard error* (SE) of the statistic. Assuming that the bootstrap distribution is relatively symmetric and bell-shaped, we create the interval using $Sample\ Statistic \pm 2 \cdot SE$. The second method uses percentiles to cut off the extreme values in the tails to leave 95% (or whatever the desired confidence level might be) of the bootstrap statistics in the middle. The boundaries for this middle 95% give the confidence interval. (There are other methods for finding confidence intervals from a bootstrap distribution but we feel these are beyond the scope of an introductory course.)

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

We see advantages in both of these methods. The $Statistic \pm 2 \cdot SE$ form of the interval helps students get ready for "formula"-based intervals, while the *percentile interval* method aids with visual understanding of what the confidence level means. For example, it is easy to see why a 99% confidence interval should be wider than a 95% interval.

Example: Figure 1 shows a bootstrap distribution of means (done in StatKey) based on a sample of prices (in $1,000's) for a brand of used car. The original sample has 25 cars with a mean price of 15.98 and standard deviation of 11.11. The percentile method interval (11.930 to 20.434) is shown on the bootstrap distribution below, while we use the standard deviation of the bootstrap statistics (2.174) to compute the interval on the right using the $Statistic \pm 2 \cdot SE$ method.

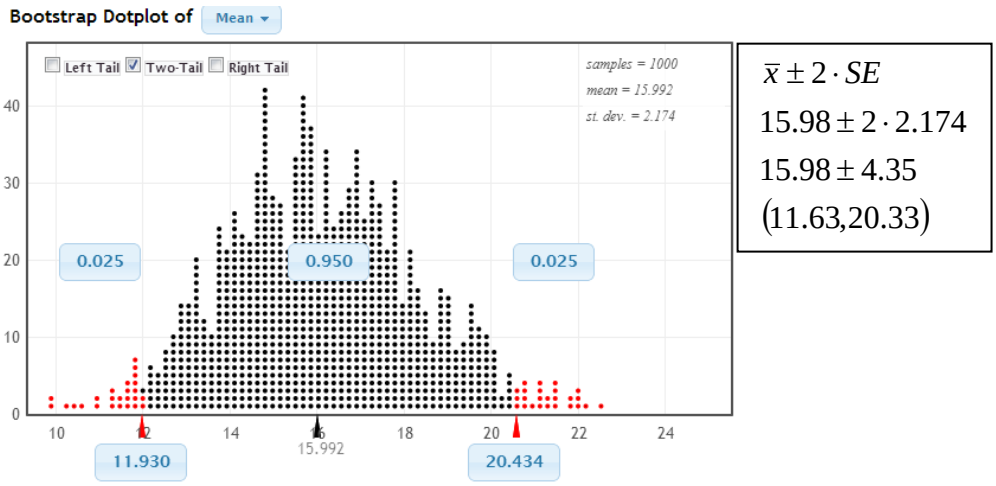


Figure 1. Bootstrap distribution and 95% CI for mean price of used cars via percentiles and SE

Note that there is some disagreement between the two methods. Indeed every different set of 1000 bootstrap samples yields a slightly different interval estimate. That is something that students (and instructors) need to get used to when dealing with simulation methods. We are no longer able to expect an exact match to the "back-of-the-book" answer to a problem. We believe this is an advantage as it reminds students that there are rarely "exact" answers in statistics. For answers that vary less from simulation to simulation, simply generate more bootstrap samples.

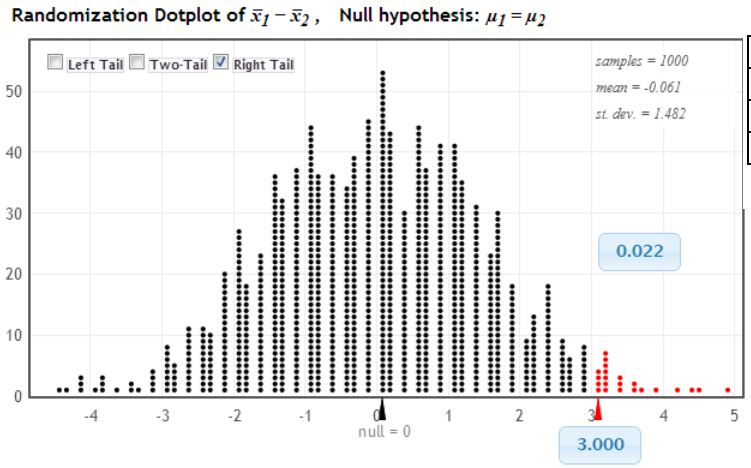
Randomization Distributions for Assessing Significance with a P-value

We start with a set of (null and alternative) hypotheses, some sample data, and a statistic measuring some aspect of that sample data. The critical question of interest is how unusual would that sample statistic be if the null hypothesis were true. A straightforward way to approach this question (and thus estimate a p-value) is to generate lots of samples showing what is likely by random chance if the null hypothesis is true, and then compute the proportion of statistics from those randomization samples that are as extreme as the statistic from the original sample.

Example: A study (Mednick et al., 2008) compared people's ability to recall a list of words after a time period in which some of the subjects had caffeine and others had a short nap (with random assignment to the groups). We wish to test whether the mean number of words recalled (μ) is higher with sleep than with caffeine, i.e. H_0: $\mu_s=\mu_c$ vs H_a: $\mu_s>\mu_c$. The study randomly assigned 12 subjects to each group, and the difference between the mean number of words recalled after sleep ($\bar{x}_s$) and after caffeine ($\bar{x}_c$) was $\bar{x}_s - \bar{x}_c = 15.25 - 12.25 = 3.0$.

How do we create the randomization samples in this situation? Since we want to know whether a difference as large as the observed sample difference is unlikely to occur by "random chance alone" and the randomness in collecting these data was due to the random assignment of subjects to the "sleep" and "caffeine" groups, we can generate new samples by randomly reassigning the 24 subjects to the two groups and assuming (under the null hypothesis of no difference in the treatments) that the number of words recalled would be the same.

Students can easily generate one or two randomization samples by hand with a physical simulation. Have them write the word counts from the original sample on 24 cards, shuffle the cards, and then deal them into two piles (one sleep, the other caffeine) of 12 values each. Finding the mean of each sample and computing the difference gives one value of a randomization statistic for this situation. In practice, we need thousands of such randomization samples to get a good estimate of the p-value, so we move quickly to technology.



	Sleep	Caffeine
subjects	12	12
mean	15.25	12.25
std. dev.	3.55	3.31

Figure 2. Randomization distribution for the difference in mean words recalled

Figure 2 shows a randomization distribution of differences in means for 1000 random reassignments of sleep and caffeine labels to the 24 word recall values. Since this is an upper-tail alternative, we estimate the p-value by finding the proportion of these samples that give a difference as large (or larger) than the observed sample difference of 3.0. This gives our estimated p-value of 22/1000 = 0.022. Since this is quite small, we have fairly strong evidence that sleep is more beneficial than caffeine in this sort of memory task. Perhaps students should avoid caffeine fueled all-nighters before a big exam and get a good night's rest instead!

WHY ARE SIMULATION METHODS VALUABLE?

Using bootstrap methods to construct confidence intervals helps students understand the general idea of a confidence interval in a relatively simple setting, and can be introduced very early in a course. The idea of cutting the tails off the distribution and keeping the middle is a powerful visual image for the students, and builds conceptual understanding without getting bogged down in formulas. The methods can easily be applied with no modifications to a wide variety of situations. Students can find a confidence interval for a slope or a standard deviation just as easily as for a mean or a proportion, since the method is exactly the same in each case. Bootstrap distributions can reinforce many of the key ideas of sampling distributions, such as variability in statistics.

The process to find a p-value in a randomization distribution is also very visually compelling and directly reinforces the definition of a p-value. Students associate finding a p-value with looking at how extreme the sample data are, in a distribution that is clearly constructed to satisfy some null hypothesis. The fact that we "assume the null hypothesis is true" is an integral part of the process, as is the idea of seeing how likely data as extreme as the sample data are to occur by random chance. The process for finding the p-value is more intuitive and concrete than looking up a standardized value in a table for some theoretical approximating distribution.

Our students are very visual learners, and many are not strong with mathematical formulas. They can see how an interval is capturing the middle 95% of values, or see how extreme the statistic is in a distribution showing what is likely if the null hypothesis is true. They can internalize that more extreme (and hence, smaller p-value) gives stronger evidence against the null hypothesis. We believe these methods make it easier for students to see the connections to the underlying concepts, while traditional formula-based methods often obfuscate those connections.

HOW SHOULD THESE METHODS BE IMPLEMENTED?

Should Simulation and Distribution-based Methods Both Be Covered, and If So, Which First?

Is it feasible, and desirable, to cover both simulation methods and traditional distribution-based methods in a single introductory course? It is feasible, since our experience is that once students understand the concepts from the simulation methods, the distribution-based methods can be covered very quickly. Whether or not it is desirable to cover both depends on the goals of the course and the student audience. If both are covered, we think the answer to the question of which should come first is fairly obvious. The background material for students to use and understand simulation-based methods is pretty minimal. Malone et al.(2010) argue for resequencing topics in an introductory course to get to the main ideas of inference much earlier. This is quite feasible to do using intuitive simulation based approaches that require little, if any, background material.

Should Intervals and Tests Both Be Covered, and If So, Which Should Come First?

We believe both should be covered, but the order question is more challenging. A number of simulation-based curriculum projects, e.g., Tintle, et al. (in press) and Zieffler et al. (2013), start with randomization tests as the initial exposure to statistical inference. However, our project (Lock et al., 2013) leads off with bootstrapping to assess the accuracy of an estimator and construct a confidence interval for a parameter. Here are a few reasons for our decision to do intervals first:

- Malone et al. (2010) point out advantages of matching the order of course topics to the order we follow in a typical data analysis. Thus we start with issues of data collection, sampling, and experimental design followed by standard numerical summaries and graphical displays. The question that naturally arises is "How accurate are those sample estimates?" This leads directly to the idea of a bootstrap as a way to assess the variability in a sample statistic.
- The question of “How accurate is my statistic?” is relatively straightforward and easy for students to comprehend, especially as compared with “How extreme would my statistic be, if the null hypothesis were true?”
- The bootstrap process is straightforward and can easily be applied to lots of different statistics. Just sample with replacement, compute the statistic of interest, and collect the results to form a bootstrap distribution. On the other hand, the methods used to create randomization samples can differ depending on the situation (See *one-crank vs. two* below).

Can We Explain Bootstrapping?

To many introductory students (and some instructors) the idea of using a bootstrap distribution from a single sample to assess the variability of a sample statistic seems a bit magical – and perhaps even looks like cheating. How can we use just the sample itself to assess how accurate the sample statistic is likely to be? We have found the following analogy to be helpful in getting students (and instructors) to see what is going on with the bootstrap process.

Think of a population as the crown of a large tree with the trunk representing a *parameter*, such as the mean being the balancing “center” of the population (See Figure 3). Imagine seeds dropped from this tree to represent statistics generated for a sampling distribution, using repeated samples from the population. Many seeds fall fairly near the trunk, but a few might drift farther away. The variability in the distances of those seeds (sample statistics) from the trunk (parameter) in this sampling distribution is a key quantity (standard error) that we need for inference.

But in statistical practice, we can’t see this whole distribution of seeds or the trunk of the tree. We only have the information from a *single* sample, just one seed, to use to try to estimate where its tree might be. What can we do with just a single seed? The answer is obvious – *grow a new tree*! That’s essentially what we are doing in creating a bootstrap distribution. We use the information in the sample, assuming that the structure (DNA) is similar to the population it came from, to generate a reasonable model for that population and then create many new samples from this new population. By observing the behavior of those new seeds (bootstrap statistics), and how far they typically fall from the new trunk (the statistic for the original sample, which we know) we can estimate how far that original seed is likely to have fallen from the trunk of its tree. As Chris Wild put it in a recent plenary talk at USCOTS 2013 “We use the bootstrap errors that we CAN see to estimate the sampling errors that we CAN’T see.”

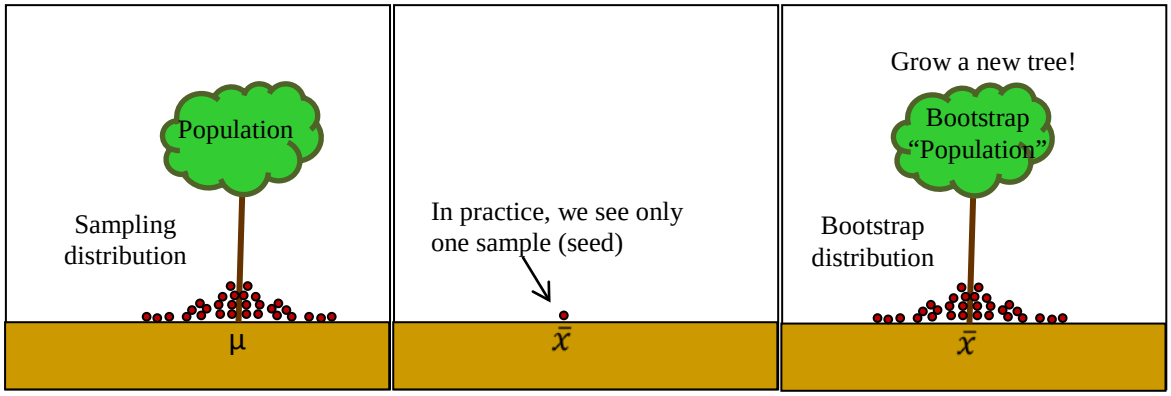


Figure 3: *Motivation for the bootstrap process using trees dropping sample seeds*

One Crank Versus Two?

John Holcomb asked a question regarding teaching randomization methods at ICOTS8 (Holcomb et al., 2010) which has become known as the "one or two cranks" debate. When testing for a relationship between two variables a natural way to simulate new samples under a null hypothesis of no association is to randomly scramble values of one of the variables to break any association with the second variable. We'll call this the *reallocation* crank and note that it is particularly appropriate for data obtained from an experiment that involved random assignment to treatment groups, like sleep and caffeine above. But suppose our data were obtained from random samples of sleepers and caffeine drinkers in a population. Another method for generating randomization samples would be to combine the data from both samples into one large group to represent the population, then get two samples with replacement from that combined sample (so both come from a "population" with the same mean) using the same sizes as the original sleep and caffeine samples. We'll call this the *resample* crank.

Should students see both of these randomization methods in an introductory course? Yes, although perhaps not in full detail. In some situations (such as a test for a single proportion), there is no feasible way to get randomization samples consistent with a null hypothesis through reallocation, thus we have to resample from a population that matches the null hypothesis. One of the nice features of randomization tests is that they allow us to make a clear connection between how randomness is used in generating the data and how it is used to simulate new samples to assess significance. A much harder question is the extent to which we should insist on students actively making this choice whenever they do a randomization test. We think that depends greatly on the level and sophistication of the course and students. The actual results will generally vary little between different randomization methods. The key point for *all* students to see is that the randomization samples are created in a way that reflects the null hypothesis.

What About Technology?

Efficient, easy to use technology is essential for having students apply simulation techniques in an introductory course. Fortunately computing power continues to increase and become more accessible. The ultimate goal is to allow students to easily use and explore simulation-based methods without being overwhelmed with technical/programming issues. Ideally, the technology should make heavy use of interactive graphics that help illustrate the main concepts of inference. For example, it should allow students to see and distinguish between the original sample, a single bootstrap or randomization sample, and the distribution of statistics from many simulated samples. Software should strike a balance between making methodologies simple while not being so automated that it becomes simply a mysterious "black box" that cranks away in the background and spits out an answer.

Here are four good sources of technology tools that are freely available, well-suited for supporting a simulation-based curriculum, and designed for introductory student use.

- *StatKey* (http://lock5stat.com/statkey/) is a set of web apps we have developed specifically to support introductory level instruction based on bootstrap intervals and randomization tests. For

further information check the help pages connected to the StatKey site or a paper specifically on StatKey (Lock Morgan et al., 2014) in the ICOTS9 proceedings.

- *Rossman/Chance Applets* (http://www.rossmanchance.com/applets/) are web apps developed by Allan Rossman and Beth Chance that include support for simulation-based activities. They do a particularly nice job of connecting physical randomizations to the computer simulations.
- *VIT: Visual Inference Tools* (https://www.stat.auckland.ac.nz/~wild/VIT/) are software modules developed by Chris Wild's group at the University of Auckland that make excellent use of real-time animations to demonstrate ideas of bootstrapping and randomization tests.
- *Mosaic* (http://mosaic-web.org/) is an R package developed by Randall Pruim, Daniel Kaplan, and Nicholas Horton that includes a number of student/instructor-friendly features for making R more accessible to students, including nice support for simulation-based methods.

Are There Textbooks/materials that Support this Approach?

In the half dozen years since George Cobb raised his challenge to move away from a Ptolemaic-focus on normal and t-based procedures, some progress is now being made on textbooks and course materials to support a simulation based approach for courses at an introductory level. Tabor and Franklin (2012) have a book geared towards secondary school students that has a sports theme, we have developed a text (Lock et al., 2013), the Catalysts for Change group at the University of Minnesota has a downloadable textbook (Zieffler et al., 2013) that uses TinkerPlots software, and Tintle et al. (in press) have a text in class test form that should be published soon.

CONCLUSION

After several years using simulation methods to introduce the key ideas of inference, we are very happy with this new approach. Students are enthusiastic about the methods, adapt to the technology easily, and show improved conceptual understanding on assessments. Formal comparisons of methods (e.g. Tintle et al., 2011) are beginning to emerge with encouraging results. We hope that we are approaching Cobb's vision and making "Now, there is no excuse" a reality.

REFERENCES

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1), Article 1.

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. In C. Reading (Ed.), *Proceedings of the 8th International Conference on Teaching Statistics*. Voorburg, the Netherlands: ISI.

Lock Morgan, K., Lock, R., Lock, P.F., Lock, E., & Lock, D. (2014). StatKey: Online tools for bootstrap intervals and randomization tests. *Proceedings of the 9th International Conference on Teaching Statistics*. Voorburg, the Netherlands: ISI.

Lock Morgan, K. (2011). *Using simulation methods to introduce inference*. CAUSE Teaching & Learning Webinar available at http://www.causeweb.org/webinar/teaching/2011-12/

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

Mednick, S., Cai, D., Kanady, J., & Drummond, S. (2008). Comparing the benefits of caffeine, naps and placebo on verbal, motor and perceptual memory. *Behavioural Brain Research, 193*, 79-86.

Malone, C., Gabrosek, J., Curtiss, P., & Race, M. (2010). Resequencing topics in an introductory applied statistics course. *The American Statistician, 64*(1), 52-58.

Tabor, J., & Franklin, C. (2012). *Statistical reasoning in rports.* New York: W.H. Freeman.

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (in press). *Introduction to statistical investigations.* Wiley.

Tintle, N., VanderStoep, J. Holmes, V. Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1). www.amstat.org/publications/jse/v19n1/tintle.pdf

Zieffler, A., & Catalysts for Change (2013). *Statistical thinking: A simulation approach to uncertainty* (second edition). Minneapolis, MN: Catalyst Press.