# STATKEY: ONLINE TOOLS FOR BOOTSTRAP INTERVALS AND RANDOMIZATION TESTS

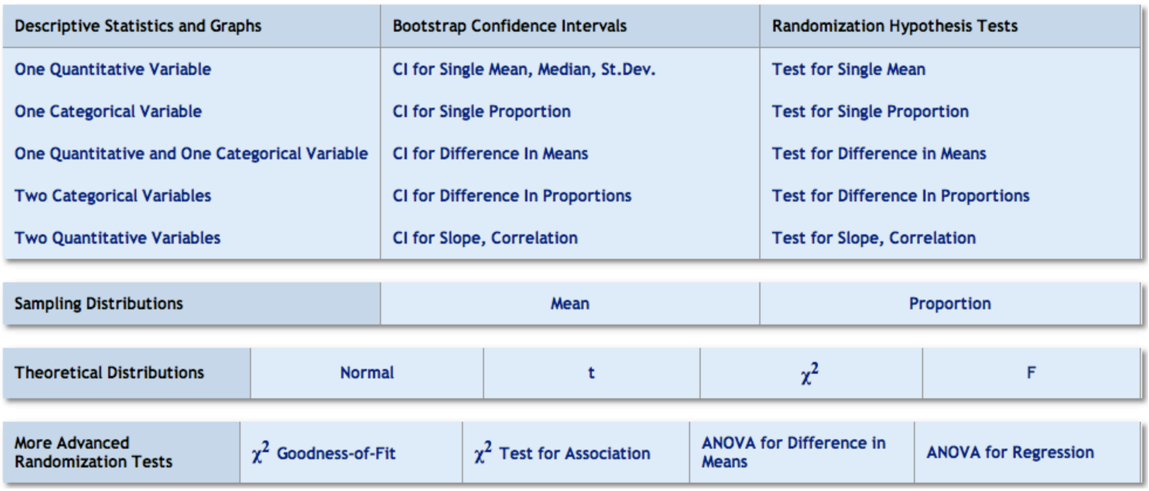
Kari Lock Morgan[1], Robin H. Lock[2], Patti Frazer Lock[2], Eric F. Lock[1], Dennis F. Lock[3]
[1]Department of Statistical Science, Duke University, Durham, NC, USA
[2]St. Lawrence University, USA
[3]Iowa State University, USA
kari@stat.duke.edu

*StatKey (www.lock5stat.com/StatKey) is free online technology created by the Lock family, designed to help introductory students understand and easily implement bootstrap intervals and randomization tests. Randomization-based methods make the fundamental concepts of statistical inference more visual and intuitive, free professors to cover inference earlier in the course, and help students see connections that are otherwise lost with numerous different formulae. To make these methods accessible to all introductory students, our goal was to create technology that is free, widely available (StatKey works in any common web browser), very easy to use, and which helps build conceptual understanding. Although particularly designed for randomization-based methods, StatKey can also be used for illustrating descriptive statistics, sampling distributions, confidence intervals, simple linear regression, and as a replacement for paper distribution tables.*

## INTRODUCTION

StatKey is a set of free online tools designed specifically for teaching bootstrap intervals and randomization tests in introductory statistics. In addition to making simulation-based methods easier to understand and implement, StatKey also provides most of the functionality you would want from software for an introductory course, such as summary statistics, data visualization, and theoretical distributions. It was designed by us (the Lock family), and implemented by computer scientists Rich Sharp, Ed Harcourt, and Kevin Angstadt. This paper is meant to be a guide to StatKey and a lens into the pedagogical reasons for its design.

Figure 1 is the homepage for StatKey, and shows the available functionalities. The top set of features is arranged with parameter type by row (One Quantitative, One Categorical, etc.), and statistical method by column (*Descriptive Statistics and Graphs*, *Bootstrap Confidence Intervals*, and *Randomization Hypothesis Tests*). Below there are three rows: *Sampling Distributions*, *Theoretical Distributions*, and *More Advanced Randomization Tests*. *Sampling distributions* illustrate the concept of sampling distributions, and connect this concept with confidence intervals. *Theoretical distributions* replace conventional distribution tables often found in the back of textbooks with more visual and easy to use online applets. *More Advanced Randomization Tests* are appropriate for chi-square and ANOVA analyses. You can return to this menu at any point by clicking on the "StatKey" box in the top left corner of every page.



| Descriptive Statistics and Graphs | Bootstrap Confidence Intervals | Randomization Hypothesis Tests |
|---|---|---|
| One Quantitative Variable | CI for Single Mean, Median, St.Dev. | Test for Single Mean |
| One Categorical Variable | CI for Single Proportion | Test for Single Proportion |
| One Quantitative and One Categorical Variable | CI for Difference In Means | Test for Difference in Means |
| Two Categorical Variables | CI for Difference In Proportions | Test for Difference In Proportions |
| Two Quantitative Variables | CI for Slope, Correlation | Test for Slope, Correlation |

| | | |
|---|---|---|
| Sampling Distributions | Mean | Proportion |

| | | | | |
|---|---|---|---|---|
| Theoretical Distributions | Normal | t | $\chi^2$ | F |

| | | | | |
|---|---|---|---|---|
| More Advanced Randomization Tests | $\chi^2$ Goodness-of-Fit | $\chi^2$ Test for Association | ANOVA for Difference in Means | ANOVA for Regression |

Figure 1. Menu for StatKey: www.lock5stat.com/StatKey

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 [© 2014 ISI/IASE]

Simulation-based methods help students understand the fundamental concepts of inference (sampling variability, confidence intervals, p-values) by using a procedure directly related to the concept at hand (Cobb, 2007). For example, consider a hypothesis test and the concept of a p-value. Traditional methods involve plugging numbers into the appropriate formula, and comparing the result to a theoretical distribution to find a p-value, offering little to no intuition for the concept of a p-value. In contrast, a randomization test involves simulating what types of statistics would be observed if the null hypothesis were true, and seeing how extreme the observed statistic is, compared with these simulated statistics. This is exactly getting at the notoriously difficult concept of a p-value that we so want our students to understand! We strongly believe in the benefits of randomization-based methods for helping students understand inference, and StatKey emerged to make these methods more easily accessible to everyone. For more information on teaching with this approach, see Lock (2014).

Although StatKey can be used with any textbook, it was designed to accompany our book, *Statistics: Unlocking the Power of Data* (Lock$^5$, 2012), and all of the built-in datasets come from this text. Clicking the dataset name in the top left corner within any StatKey procedure opens a drop-down menu of these datasets. You can also import your own data into StatKey by choosing “Edit Data” and copy/pasting from a spreadsheet, other electronic source, or typing values directly, taking care to match the format that StatKey expects. Categorical information for proportions can also be entered directly as counts. See the help menu in StatKey for more detailed instructions.

StatKey is available at lock5stat.com/statkey, and works on any common web browser. StatKey is also available freely as a Google Chrome app, which allows it to be used even without an internet connection. We encourage you to visit StatKey now, and click along!

## BOOTSTRAP CONFIDENCE INTERVALS

Click on “CI for Single Mean, Median, Std. Dev.,” and a new simulation page opens. Everything in blue is clickable. The default dataset is “Ottawa Senators (penalty minutes)”, click here for a drop-down menu of other datasets, and choose “Florida Lakes (Mercury in Fish),”. data on average mercury level of fish (large mouth bass) for 53 lakes in Florida (Lange et al., 2004).

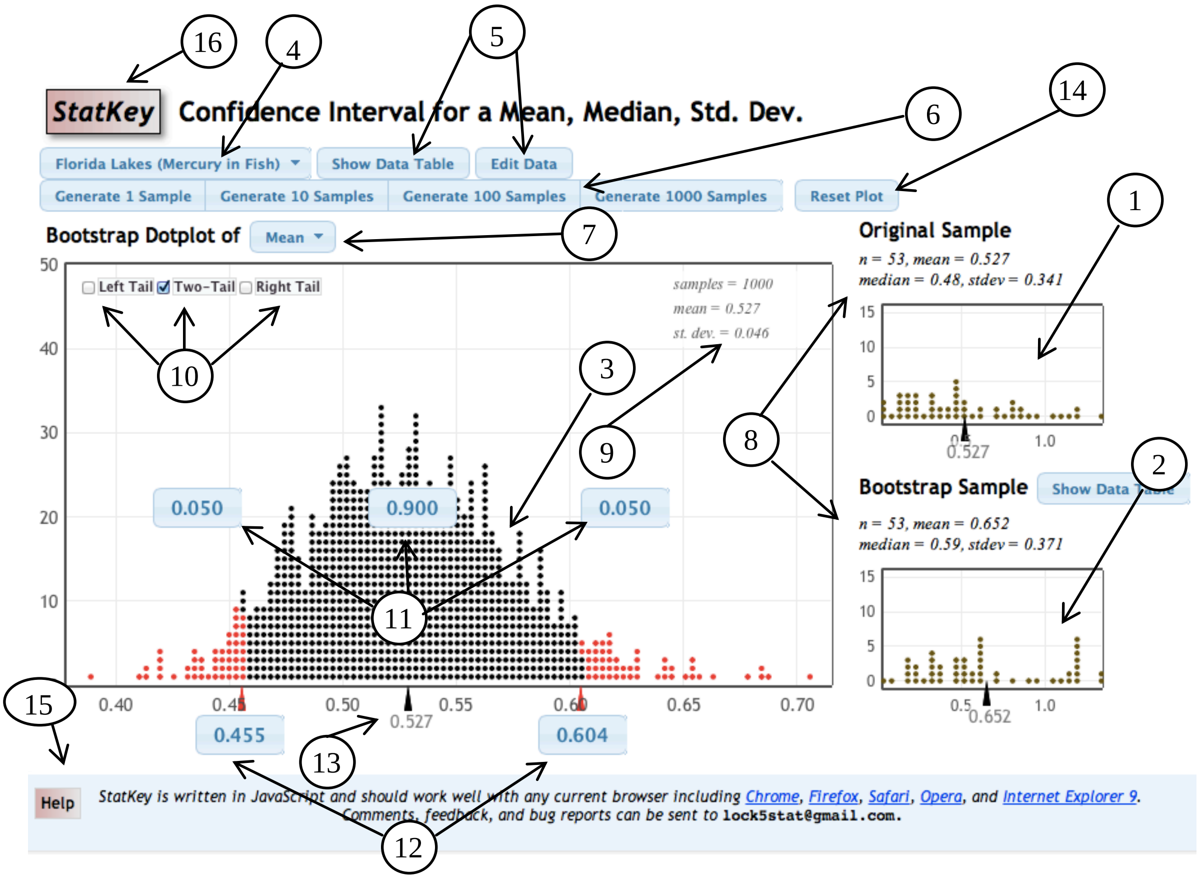
The relevant summary statistics and visualization for the data appear below *Original Sample*. For this quantitative variable, we see the sample size, mean, median, standard deviation, and a dotplot. We start with this to encourage students to first look at the summary statistics and plot of the sample data, *before* doing inference. This is shown in the top right of Figure 2. The sample mean is 0.527 ppm (the FDA action level in the USA is 1 ppm, in Canada the limit is 0.5 ppm). How much might this mean vary from sample to sample? Let’s bootstrap to find out!

### *Generating a Bootstrap Distribution*

By clicking on “Generate 1 Sample,” we generate one bootstrap sample, a sample of the same size as the original sample selected by sampling from the original sample with replacement. The summary statistics and visualization of this bootstrap sample are displayed under “Bootstrap Sample.” This is a nice place to stop and ensure that students actually understand the process of bootstrapping. It can help to focus on particular units in this discussion, for example, in the particular bootstrap sample shown in Figure 2, the lake with the highest mercury level happened to be sampled twice, and the lake with the lowest mercury level was not sampled at all. By moving your cursor over a single dot in the dotplot, you can see the actual data value, which can be helpful for this discussion.

The bootstrap sample mean of 0.629 ppm is higher than the actual sample mean. This value appears under “Bootstrap Sample” and also as one dot in the bootstrap distribution. Before simulating more samples, this is another good point to stop and make sure students see the connection. The idea that each dot in the bootstrap distribution is one *statistic* from an entire bootstrap sample, not a single data point, can be difficult for students to grasp and is worth spending time on. We think the ability to generate just one sample at a time is important for helping students see the connection between the statistic under “Bootstrap Sample” and the corresponding dot in the bootstrap distribution. Also, moving the cursor over any dot in the bootstrap distribution will display the bootstrap sample that yielded that bootstrap statistic, another feature that facilitates understanding.

Once students understand how a bootstrap sample is achieved, and what each dot in the bootstrap distribution represents, we can click on "Generate 1000 Samples," which gives us a bootstrap distribution. We can click this repeatedly for more simulations (intervals and p-values get more precise as the number of simulated samples increases). This distribution is shown in Figure 2. We like bootstrap distributions because they focus on a key idea of inference, how much statistics vary from sample to sample, in a way that is more intuitive and meaningful than formulas.



samples = 1000
mean = 0.527
st. dev. = 0.046

n = 53, mean = 0.527
median = 0.48, stdev = 0.341

n = 53, mean = 0.652
median = 0.59, stdev = 0.371

StatKey is written in JavaScript and should work well with any current browser including Chrome, Firefox, Safari, Opera, and Internet Explorer 9. Comments, feedback, and bug reports can be sent to lock5stat@gmail.com.

Figure 2: Bootstrap confidence interval for a mean.

Each of the numbered arrows in Figure 2 is described below:

1. Display of the original sample.
2. Display of a particular bootstrap or randomization sample.
3. Dotplot of bootstrap/randomization statistics. Mouse over any dot to see the sample that produced it!
4. Choose one of the built-in datasets from the text.
5. Display or edit a table of the data. If importing your own data, paste it in here.
6. Generate 1, 10, 100, or 1000 samples at a time.
7. Change the statistic, null hypothesis, or randomization method (when enabled).
8. Summary statistics for the original sample and a particular bootstrap/randomization sample.
9. Summary statistics for the bootstrap/randomization distribution.
10. Select regions in either or both tails of the bootstrap/randomization distribution. Two-tail gives equal proportions.
11. Editable values for proportions in different regions of the bootstrap/randomization distribution.
12. Editable endpoints for tail regions of the bootstrap/randomization distribution.
13. Mean for a bootstrap distribution, null parameter value for a randomization distribution.
14. Start over with a new simulation.
15. Get help (including videos) on StatKey features.
16. Return to the main StatKey menu

*Finding a Confidence Interval from a Bootstrap Distribution*

One way of creating a 95% confidence interval from a bootstrap distribution, if the distribution is approximately bell-shaped, is to use *statistic* ± 2·*SE*, with the standard error estimated as the standard deviation of the bootstrap distribution. This facilitates the transition to normal and t-based methods, and also helps students understand the meaning of the standard error (another potentially difficult concept). The summary statistics for the bootstrap distribution are given in the top right corner of the dotplot, and from Figure 2 we see the standard deviation of the bootstrap statistics for the mercury example is 0.046 (note that this is very close to the theoretical $s/\sqrt{n} = 0.341/\sqrt{53} = 0.0468$, and much more intuitive). Therefore, a 95% confidence interval is *statistic* ± 2·*SE* = 0.527 ± 2(0.046) = (0.435, 0.619). We are 95% confident that the average mercury level of large mouth bass in Florida lakes is between 0.435 and 0.619 ppm.

We can also create a confidence interval via the percentiles of the bootstrap distribution; if the bootstrap distribution is roughly symmetric, an approximate C% confidence interval contains the middle C% of bootstrap statistics. This formula-free approach helps build intuition for the meaning of a confidence level. Click the checkbox next to "two-tail" to bring up several editable blue boxes, including the proportion in the middle of the distribution and in each of the tails. By default, 95% of the bootstrap statistics are contained in the middle, so the two corresponding endpoints give a 95% confidence interval. For a 90% interval, we simply click on the box in the middle and change 0.95 to 0.90. The endpoints adjust to contain the middle 90% of bootstrap statistics, as shown in Figure 2, giving a 90% confidence interval of 0.455 to 0.604 ppm. We feel it is pedagogical advantage that students have to know how to interact with the bootstrap distribution.

## RANDOMIZATION HYPOTHESIS TESTS

Let's click on "Test for Difference in Means" from the main menu, and choose the dataset "Mindset Matters (WgtChange by Informed)". This dataset is from a study (Crum and Langer, 2007) in which hotel maids were randomly divided into two groups; one group was informed that the work they do satisfies the surgeon general's recommendations for an active lifestyle (which it does), and the other group was not informed of this. The variable *Informed* is whether the information was given, and *WgtChange* represents weight change over four weeks. In the original sample, we see that the informed maids lost 1.59 more pounds, on average, than the non-informed. Is this statistically significant? Let's conduct a randomization test to find out!

*Generating a Randomization Distribution*

Generating a randomization distribution in StatKey is similar to generating a bootstrap distribution, except that randomization samples are simulated in a way that is consistent with the null hypothesis.

The null hypothesis for the Mindset Matters experiment is that *Informed* status has no effect on weight gain, so each maid would have gained the same amount regardless of the group she was assigned to. We can simulate other potential datasets we could have obtained by random chance if the null hypothesis were true. The "random chance" is the random assignment to either be informed or not, so we reallocate the maids to the two groups, keeping their response values (weight change) fixed. Click "Generate 1 Sample" to do this, and the results are displayed under Randomization Sample. Again, this process is good to discuss with students. It can help to focus on a single unit (outliers are easy to spot), and watch that particular value to see which group it is randomized to. As with bootstrap distributions, each randomization statistic is plotted with a dot in the randomization distribution, and once the process is understood, we can generate thousands of samples to create a randomization distribution. This distribution is shown in Figure 3.

Analyzing a randomized experiment is the most straightforward way in which randomization samples are created, because units are reallocated to treatment groups, mimicking the actual experiment. However, this is not the only way to create randomization samples under the null hypothesis. For a single proportion, we mimic coin flips under a specified null probability. For a single mean, we add a constant to all sample values to satisfy the null, then we bootstrap from this shifted sample. For a correlation, one of the variables is randomly scrambled, breaking any association. For observational studies, rather than reallocating units to groups, we may want to instead shift the groups to have the same mean, or else combine them into one group, and bootstrap. By default StatKey uses reallocate to create randomization samples for two groups, but this may be changed by clicking on "reallocate" (in practice, the p-values are usually very close regardless of which randomization method is used). Individual instructors may choose to leave out much of this detail. The key point to emphasize is that a randomization sample is a sample simulated as if the null hypothesis were true.

*Finding a P-value from a Randomization Distribution*

We want to see how extreme our observed statistic is relative to the randomization distribution. We click the check box next to either left tail, right tail, or two-tail, depending on whether we are doing a one or two-tailed test. For the Mindset Matters example we do a right-tailed test because the study was designed to see if the information promotes weight loss. This brings up an editable blue box on the distribution for the proportion of statistics in the tail (0.025 by default), and an editable box on the horizontal axis corresponding to the endpoint for this tail probability. We change the endpoint to match the observed statistic (1.59), and the upper tail proportion changes accordingly. As shown in Figure 3, the proportion in the upper tail is 0.005; 5 out of the 1000 simulated statistics are as extreme as the observed statistic, and are colored red on the dotplot. This provides a p-value for the test: the chance of getting a statistic as extreme as that observed, if the null hypothesis is true, is approximately 0.005. To reword this in the context of the study: if the information had no effect on weight change, we would see results as extreme as those observed in about 5 out of 1000 randomizations. We have statistically significant evidence that informing maids that the work they do satisfies the Surgeon General's recommendations for an active lifestyle causes them to lose more weight, on average, than maids not given this information.

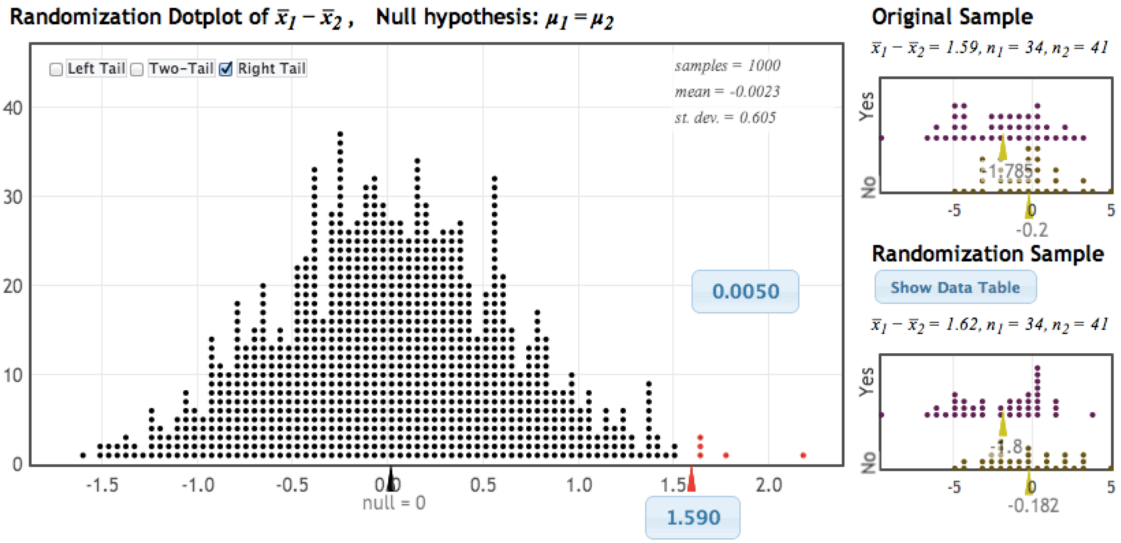

Randomization Dotplot of $\bar{x}_1 - \bar{x}_2$, Null hypothesis: $\mu_1 = \mu_2$
□ Left Tail □ Two-Tail ☑ Right Tail
samples = 1000
mean = -0.0023
st. dev. = 0.605
40
30
20
10
0
0.0050
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
null = 0
1.590
Original Sample
$\bar{x}_1 - \bar{x}_2 = 1.59, n_1 = 34, n_2 = 41$
Yes
No
-1.785
-5
0
5
-0.2
Randomization Sample
Show Data Table
$\bar{x}_1 - \bar{x}_2 = 1.62, n_1 = 34, n_2 = 41$
Yes
No
-1.8
-5
0
5
-0.182

Figure 3: Randomization test for a difference in means.

We love how this approach makes a p-value so straightforward, visual, and intuitive! We particularly like the fact that students have to know how to interact with the randomization distribution; the answer isn't just given to them. It would have been easy to program StatKey to automatically adjust the endpoint to match the observed statistic, and automatically give the p-value, but we made a conscious decision *not* to do this. We believe there is real pedagogical value in making students think about what they want (while also making it very easy for them to get what they want). The question "How extreme is my observed statistic?" should be at the forefront of students' minds when doing a randomization test, and the design of StatKey ensures that this is so.

*Advanced Randomization Tests: Chi-Square and Analysis of Variance (ANOVA)*

StatKey also provides the option to do more advanced randomization tests, such as chi-square tests for goodness-of-fit and association, and ANOVA for difference in means and regression. The format for these randomization tests matches that shown in Figure 3, except using either the $\chi^2$-statistic or $F$-statistic. For the original sample and each randomization sample, the details for calculating the statistic (observed and expected counts or the full ANOVA table) are displayed with "Show Details". The randomization distribution for either of these statistics can be used as a standalone test, or as a nice way to motivate the corresponding theoretical distributions.

## OTHER FEATURES OF STATKEY

"Descriptive Statistics and Graphs" provides relevant summary statistics and visualization for one or two variables. For example, for one quantitative variable, the sample size, mean, standard deviation, and five number summary are given, and students can choose between viewing the data as a dotplot, histogram, or boxplot. For the histogram a slider controls the number of bins.

"Sampling Distributions" allows simulation of sampling distributions for means or proportions. The format matches that of bootstrap intervals, except the data are population data, the sample size can be specified, and samples are drawn without replacement. While this is not particularly useful for analyzing data, it is great for helping students understand sampling distributions, and for illustrating the central limit theorem. There is also an option to show confidence intervals, in which a confidence interval is displayed as a horizontal bar for each sample generated, and coverage rate can be tracked. This is useful for illustrating that that intervals vary from sample to sample. Also, moving the cursor over a dot in the sampling distribution highlights the corresponding interval (and vice versa), demonstrating the connection between extreme statistics in the tails of the sampling distribution and intervals that miss the parameter.

"Theoretical Distributions" replaces the cumbersome, difficult to understand, unintuitive paper tables found in the back of many textbooks with a visual, easy to use electronic version. The format matches that used for bootstrap and randomization distributions, with blue editable boxes on the distribution for tail or center proportions and on the horizontal axis for endpoints. The only difference is that rather than a simulated dotplot, students see a smooth theoretical distribution. We hope that this parallelism helps ease the transition from simulation to distribution-based methods.

## CONCLUSION

StatKey is free web-based technology, available to anyone, designed with conceptual understanding and ease-of-use in mind. Below we highlight a few key pedagogical features:

- Ability to simulate one to many samples
- Clear distinction between the original sample, a single simulated sample, and the distribution of simulated statistics, with all three shown consistently and simultaneously.
- Students interact with the bootstrap/randomization distribution; some thought required!
- Consistent interface for bootstrap intervals, randomization tests, theoretical distributions.

With StatKey, we hope that technology will no longer be a limiting factor for those wishing to teach simulation-based methods. Whether it's as a one-time demo in class, as a replacement for paper tables, as supplemental technology for teaching simulation-based methods, or as standalone technology for your course, we hope you give StatKey a try!

## REFERENCES

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

Crum, A., & Langer, E., (2007). Mind-set matters: Exercise and the placebo effect. *Psychological Science, 18*, 165-171.

Lange, T., Royals, H., & Connor, L. (2004). Mercury accumulation in largemouth bass (Micropterus salmoides) in a Florida Lake. *Archives of Environmental Contamination and Toxicology, 27*(4), 466-471.

Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., Lock, D. F. (2013). *Statistics: Unlocking the power of data* (1st edition). Hoboken, NJ: John Wiley & Sons.

Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., Lock, D. F. (2014). *Intuitive introduction to the important ideas of inference*. Paper presented at the Ninth International Conference on Teaching Statistics (*ICOTS 9*), Flagstaff AZ, USA.