FROM DATA TO DECISION-MAKING: USING SIMULATION AND RESAMPLING METHODS TO TEACH INFERENTIAL CONCEPTS

Mia L. Stephens[1], Robert H. Carver[2] and Don McCormack[1]
[1]JMP Software, Cary NC, USA
[2]Stonehill College & Brandeis University, Massachusetts, USA
mia.stephens@jmp.com

It is the year 2014. Despite advances in technology and the availability of interactive software and teaching tools, it is still the norm to use normality-based methods, and even look-up tables, to teach statistical inference. This puts a heavy burden on our students, who must struggle through difficult theory before they're taught how to make decisions and draw inferences from data. Even then, do they understand what a p-value is or what a confidence interval represents? Is there a better way? Interactive computer simulations and resampling methods can help bridge the gap between graphs and summary statistics and inference, providing a gentler and more natural transition. Until recently, these methods required add-ins, specialized programs or custom code. Today, these techniques are available in mainstream statistical software. In this talk, we illustrate how to use simulations, bootstrapping, and randomization tests in JMP® to introduce sampling distributions and explore core inferential concepts.

INTRODUCTION

It is well documented that simulations and resampling methods are effective tools for teaching statistical concepts (Wood 2005, Mills 2002). The GAISE report (Aliaga et al, 2012), which provides guidelines for instruction in statistical education, recommends using technology to "perform simulations to illustrate abstract concepts." With recent advances in computing power and expanded availability of technology, the use of computer simulation is becoming more common, and increasingly more popular, in all levels of statistics education. Today, a Google search will yield hundreds of hits on the topics of simulation, bootstrapping, and randomization tests – from articles and blogs, to sites dedicated to simulation applets and programs. Yet, as George Cobb (2007) wrote in the first issue of the journal *Technology Innovations in Statistical Education,*

> What we teach is largely the technical machinery of numerical approximations based on the normal distribution and its many subsidiary cogs. This machinery was once necessary, because the conceptually simpler alternative based on permutations was computationally beyond our reach. 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.

This begs the question, why have many educators yet to integrate simulation and randomization methods into the classroom? One potential reason is that few textbooks (or e-books) have fully integrated randomization-based methods as a pedagogical approach, relying heavily on traditional practices (Cobb, 2013). Some recent exceptions are Tintle et al (2013), Lock et al (2013), and Good (2013), which are built around using computer simulation and resampling methods to introduce statistical inference. Lack of consistent access to computers, connection to the internet, or technical savvy to understand and utilize the methods effectively might be additional obstacles. Another potential issue may be related to the sheer number of applets, add-ins and simulators available, and the disconnect between these tools for developing statistical reasoning skills and the statistical tools required for analyzing data.

ALL METHODS, ONE PROGRAM

Statistics instructors have many objectives to satisfy in a narrow timeframe. While simulation and randomization-based methods can be more effective in introducing statistical concepts than traditional methods, getting students up to speed on accessing, using and interpreting the results from add-ins, applets or other simulation programs can be time consuming. Since these

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

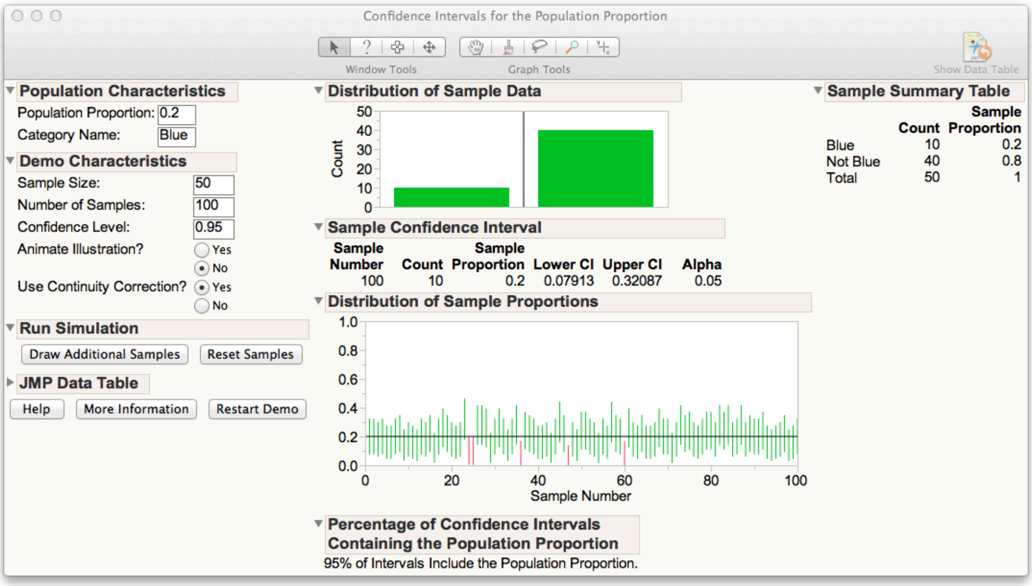
tools generally don’t have the capacity for true data analysis, additional statistical software is often needed, requiring additional time and valuable resources. Whereas statistical software can become a resume item and hiring point for students entering the workforce, few employers will recognize specialized add-ins or simulation applets as viable statistical tools.

In this paper, we introduce a statistical tool for data analysis and exploration, JMP® 11 Pro (pronounced Jump), which can also be used for simulation and randomization-based methods. JMP® is an interactive and dynamic tool for statistical analysis and discovery, which has a rich programming language for building custom applications, add-ins and simulations. One-click bootstrapping and Monte Carlo simulations are built-in features in JMP® Pro, and a number of simulations are built into the product as sample scripts or add-ins. In addition, a comprehensive suite of freely available add-ins for educational simulations and randomizations tests can be installed and accessed through a JMP menu.

SIMULATIONS

A simulation provides an imitation of reality. Simulations in the statistical context provide an understanding of the behavior of a chosen statistic under a given set of conditions, after repeated iterations. Computer simulations facilitate the learning experience, allowing the student to individually explore and develop an understanding of difficult or abstract concepts (Mills, 2002).

A number of simulations for exploring statistical concepts are available directly in JMP®. In one program, a student can, for example, simulate data to explore the central limit theorem or tosses of a fair coin, and use built in visualization and analysis tools to explore the results of the simulations. A comprehensive collection of simulations, Concept Discovery Modules, are also available as an add-in, providing options for interactively exploring sampling distributions, confidence intervals, hypothesis tests, probability distributions, regression, and t-tests and ANOVA. An example, the Confidence Intervals for Population Proportions module, is shown in Figure 1.



Population Characteristics

Population Proportion: 0.2

Category Name: Blue

Demo Characteristics

Sample Size: 50

Number of Samples: 100

Confidence Level: 0.95

Animate Illustration? ○ Yes ● No

Use Continuity Correction? ● Yes ○ No

Sample Confidence Interval

Sample Number	Count	Sample Proportion	Lower CI	Upper CI	Alpha
100	10	0.2	0.07913	0.32087	0.05

Sample Summary Table

	Count	Sample Proportion
Blue	10	0.2
Not Blue	40	0.8
Total	50	1

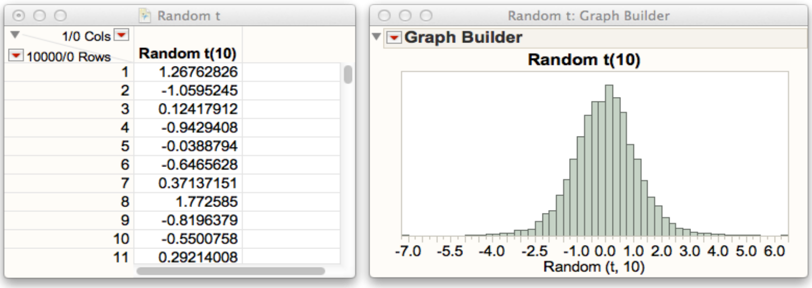
Percentage of Confidence Intervals Containing the Population Proportion

95% of Intervals Include the Population Proportion.

Figure 1 Confidence Intervals for Population Proportions module

For ultimate flexibility, the JMP formula editor and JSL, the JMP scripting language, can be used to create custom simulations using built-in functions. In Figure 2, the Random t function was used in a data table column formula to simulate 10,000 observations from the t distribution with 10 degrees of freedom, and results were graphed using the drag-and-drop Graph Builder®.

```
Random t(10);
```



Random t

1/0 Cols
10000/0 Rows

	Random t(10)
1	1.26762826
2	-1.0595245
3	0.12417912
4	-0.9429408
5	-0.0388794
6	-0.6465628
7	0.37137151
8	1.772585
9	-0.8196379
10	-0.5500758
11	0.29214008

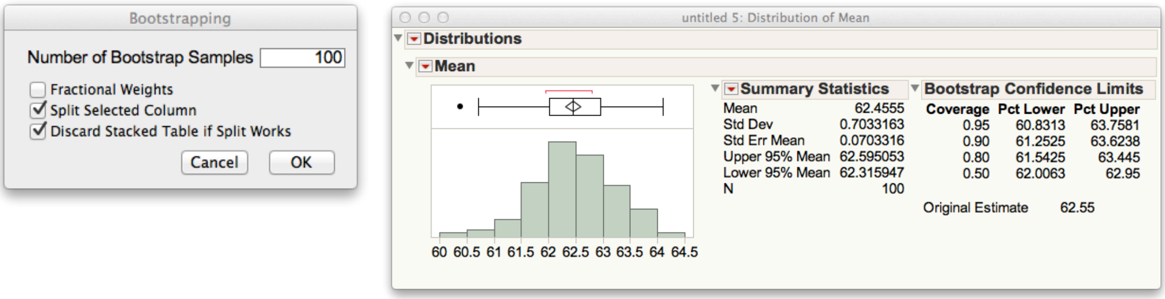
Figure 2 Random t function table of simulated data and graph in Graph Builder

BOOTSTRAP CONFIDENCE INTERVALS

The theory behind resampling methods goes back to the 1930’s and Sir R.A. Fisher (1971). Bootstrap resampling, proposed by Bradley Efron (1979), involves sampling (bootstrapping) from a data set, with replacement, in order to form a new data set of the same size. Repeatedly sampling from the original sample of data approximates the process of taking repeated samples from the target population of interest. The result is a large set of bootstrapped samples, which can be used to estimate the standard error and confidence intervals for the chosen statistic.

Since bootstrapping require minimal assumptions and can be used to estimate the sampling distribution of nearly all statistics, using bootstrapped methods to teach inference provides an intuitive and more direct educational progression (Lock, 2013). Students move from learning about the sample and sample statistics directly into exploring inferences about the population, allowing the more theoretical discussions of normal and t-based inference and distributional assumptions to be delayed until the student has a grasp of the fundamental concepts.

Bootstrapping and bootstrapped percentile confidence intervals are directly available in JMP® 11 Pro, and bootstrapped samples can be generated for most statistics. Right-clicking on statistical output launches a simple and intuitive interface (on the left in Figure 3), and results can be interactively explored and summarized using built-in analysis features (right, Figure 3).



Bootstrapping

Number of Bootstrap Samples: 100

□ Fractional Weights
☑ Split Selected Column
☑ Discard Stacked Table if Split Works

untitled 5: Distribution of Mean

Distributions

Mean

Summary Statistics

Mean	62.4555
Std Dev	0.7033163
Std Err Mean	0.0703316
Upper 95% Mean	62.595053
Lower 95% Mean	62.315947
N	100

Bootstrap Confidence Limits

Coverage	Pct Lower	Pct Upper
0.95	60.8313	63.7581
0.90	61.2525	63.6238
0.80	61.5425	63.445
0.50	62.0063	62.95

Original Estimate 62.55

Figure 3 Bootstrapping interface (left) and summary of analysis results (right)

RANDOMIZATION TESTS

Randomization tests involve simulating the distribution of sample statistics (or test statistics) one would observe under the null hypothesis and then comparing the observed statistic to this simulated distribution, providing an empirical estimate of the p-value. Lock et al (2013), who use randomization tests to introduce concepts of hypothesis testing, provide the following rationale for this approach,

> We use ideas such as randomization tests and bootstrap intervals to introduce the fundamental ideas of statistical inference. These methods are surprisingly intuitive to novice students and, with proper use of computer support, are accessible at very early stages of a course…. We believe that this approach helps students realize that although the formulae may take different forms for different types of data, the conceptual framework underlying most statistical methods remains the same.

In JMP® 11, randomization tests for a variety of hypothesis tests involving proportions and means are available in a new freely-available add-in. The user specifies the type of test, the number of samples, the randomization method, the data set and the variables. Results for the randomization samples, the original sample and the selected randomization sample are displayed in separate graphs in one interactive window.

CONCLUSION

Computer simulation and technology are enabling tools, allowing statistics educators to move beyond and perhaps abandon classical methods of instruction. As George Cobb (2007) wrote in the same issue of the journal *Technology Innovations in Statistical Education*,

> Just as computers have freed us to analyze real data sets, with more emphasis on interpretation and less on how to crunch numbers, computers have freed us to simplify our curriculum, so that we can put more emphasis on core ideas like randomized data production and the link between randomization and inference, less emphasis on cogs in the mechanism, such as whether 30 is an adequate sample size for using the normal approximation.

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, R., Utts, J., Velleman, P., & Witmer, J. (2012). *Guidelines for Assessment and Instruction in Statistics Education: College Report.* American Statistical Association. Retrieved January 7, 2014 from http://www.amstat.org/education/gaise/

Cobb, G. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum. *Technology Innovations in Statistics Education, 1*(1), Article 1. Retrieved January 7, 2014 from http://escholarship.org/uc/item/6hb3k0nz#page-1

Cobb, G. (2013). Comment: Technology and the Future of Statistics Education. *Technology Innovations in Statistical Education, 7*(3). Retrieved January 7, 2014 from http://escholarship.org/uc/item/1j7116jx#page-6

Efron, B. (1979). Bootstrap Methods: Another look at the jackknife. *The Annals of Statistics, 7*(1).

Fisher, R. A. (1971), *The Design of Experiments* (9th ed.). Macmillan. (Originally published in 1935, New York: Hafner.)

Good, P. (2013), *Introduction to Statistics Through Resampling Methods and R* (2nd ed.). Wiley.

JMP®, Version 11 Pro, SAS Institute Inc., Cary, NC, 1989–2013.

Lock R., Lock, P., Lock, K. L., Lock, E., & Lock, D. (2013), *Statistics, Unlocking the Power of Data*. Wiley.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: a review of the literature. *Journal of Statistics Education, 10*(1). Retrieved January 7, 2014 from www.amstat.org/publications/jse/v10n1/mills.html

Tintle, N,. Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (forthcoming). *Introduction to Statistical Investigations.* Wiley.

Wood, M. (2005). The Role of Simulation Approaches in Statistics. *Journal of Statistics Education, 13*(3). Retrieved January 7, 2014 from www.amstat.org/publications/jse/v13n3/wood.html