TEACHING RESAMPLING IN AN INTRODUCTORY STATISTICS COURSE

Webster West
Department of Statistics,
North Carolina State University, Raleigh, NC, USA
websterwest@ncsu.edu

There is a movement underway among statistical educators to use resampling techniques in introductory statistics courses. For these techniques to be effective, students must have access to appropriate computer software. To help in this process, a number of resampling applets are now available within the StatCrunch package, which can be customized using specific data sets. These applets are discussed in detail in the context of classroom activities with an emphasis on how they may be used to enhance the learning process. Some practical issues related to introducing resampling methods into an introductory course are also discussed.

INTRODUCTION

At the inaugural United States Conference on Teaching Statistics (USCOTS) in 2005, George Cobb created a firestorm among statistical educators with a very convincing argument for replacing much of the standard introductory statistics curriculum based on normal distribution theory with a resampling approach. In his address and subsequent paper (Cobb, 2007), his argument compared the rather convoluted path through basic inference found in the standard curriculum with the overly complicated model of the universe put forward by Ptolemy. As Ptolemy added an ever-increasing number of orbital paths to explain the observed movements of planets, Cobb noted the standard curriculum is full of complicated bifurcation points inserted in order to overcome issues related to the necessary assumptions required for standard inferential methods. Cobb pointed out that technology had advanced to the point that alternative computationally intensive approaches were now plausible for use at the introductory level. These methods, he argued, require far less in the way of assumptions and might allow students to focus on the big picture concepts rather than on the gritty details with the standard approach.

As was the case with many in attendance, I left Cobb's address energized to explore these ideas for teaching introductory statistics. My interest was spurred not only because of my role as a statistics educator but also because of my role as developer of the StatCrunch software package, which has a strong presence in statistical education. I began focusing on ways to present resampling ideas in my own classes and on developing the corresponding software to support this implementation. On the content side, I collaborated with Roger Woodard on the INCIST (Improving National acceptance of Computationally Intensive Statistical Techniques) project funded by the National Science Foundation (DUE #0817262). This project developed a number of teaching techniques for incorporating resampling into introductory statistics courses using several activities where students were asked to initially collect resamples by hand before being introduced to software. These ideas were presented at a number of teacher training workshops offered around the United States.

On the software side, I began by adding the capability within StatCrunch to easily collect multiple samples from columns of data and to compute and store basic statistics based on these samples. This preliminary work allowed for a very manual method of implementing basic resampling methods using the software. One drawback to this approach is that it required the storage of large amounts of data in the data table which could be quite memory intensive. As a follow up, I developed a general resampling tool for the package that allowed users to implement very sophisticated resampling ideas provided they understood how to specify rather complicated mathematical expressions. Realizing some of the potential issues with using either of these approaches with students in the introductory classroom, I then focused on building interactive applets into StatCrunch that not only allowed for doing the necessary resampling calculations but also provided a visualization of how the methods worked. The applets were designed to be complimentary to the INCIST activities so that teacher's could first build the intuitive ideas behind resampling methods using the activities. These activities were then well connected to the applets for repeating the calculations a much larger number of times. I have used this merger between the

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

activities and the applets for a number of semesters in my own courses. In the remainder of this manuscript, I will outline one of the examples of this implementation and follow up with a brief discussion of my initial observations on using resampling in introductory statistics courses.

RANDOMIZATION EXAMPLE

Perhaps the penultimate topic covered in most introductory statistics courses is the two-sample *t* test for comparing two means. Indeed, such an example served as the foundation for Cobb's keynote address. Cobb's suggested a randomization test as a replacement for this standard procedure. The idea being that student's could more readily understand the notion of what type of results might happen due to chance by randomly assigning data values to two groups and computing the difference in means between the groups. This rather intuitive concept can be introduced at almost any point in the typical semester without the need for much in the way of background. This randomization approach certainly does not require much in the way of distribution theory, which takes up much of the standard curriculum.

The first activity developed as part of the INCIST project applies the randomization approach to a natural setting where two groups of ten students are randomly assigned to two different version of an exam, one yellow and one white. After the exam, both groups of students complain that their version of the exam was more difficult. The observed difference between the two means (yellow – white) was 78.3 – 71 = 6.3. The raw data is available at http://www.statcrunch.com/app/index.php?dataid=716210 in both stacked and split formats. The question then becomes is there significant evidence that the typical student will score lower on one version of the exam than the other or is this sort of difference likely to occur simply due to the random assignment of the students to the two versions of the exams. To better understand the types of differences that occur due to chance, one could simply mimic the original assignment of students to exams by randomly allocating the scores to the two different versions of the exam and computing the difference between the two means. With this randomization approach, one is saying that the version of the exam has no impact on a student's score. This allows one to isolate the impact of the random chance assignment on the resulting difference between means.

For students to better grasp the randomization approach, they are asked to take part in an activity where pairs of students carry out a single randomization step by hand. Each pair reports the difference between the two means when the scores are randomly assigned to the yellow and white groups. To accomplish the task, students are given a stack of 20 notecards containing the original 20 exam scores. The student then shuffles the cards and deals them into two groups, one representing yellow and the other representing white. This process is equivalent to the notion that a student will get the same score regardless of the exam they take (no exam effect). The student then averages the 10 scores in each of the groups and computes the difference between the two means.

Each student then records their difference in means on a sticky note and places the note at the proper location on a number line drawn on the board at the front of the room. The students are encouraged to focus on what type of difference between the two means is likely to occur due to random chance. Students with a difference that is larger in absolute value than 6.3 (less than -6.3 or greater than 6.3) use a different color of sticky note. Using the different color, allows the class to easily estimate the proportion of times the difference is larger in magnitude than the observed value. After all of the students have added their randomization results to the mix, the instructor leads the class in a discussion of whether or not the original observed difference of 6.3 seems strange compared to the differences from the randomizations. For this data, roughly one in four of the differences will be larger in magnitude than the observed difference of 6.3. For smaller classes, with relatively few randomizations reported, the discussion will generally not be overly conclusive. This leads nicely to the idea of using technology to repeat this process a few thousand more times to better estimate the P-value from the randomization approach applied to this scenario. The term "P-value" is used rather loosely in this situation and it need not be mentioned in class if formal hypothesis testing jargon has not yet been introduced. Instead, the focus may be put simply on the idea of evaluating the likelihood of a value as or more extreme than the observed difference occurring due to random chance. The same is true for terms like "null hypothesis" and "alternative hypothesis". These terms are not necessarily required in the discussion of this example.

A screen shot of the dialog window used to create the follow up StatCrunch randomization applet is shown in Figure 1. This dialog can be produced in StatCrunch by choosing the *StatCrunch* > *Applets* > *Resampling* > *Randomization test for two means* menu option. The inputs shown in the window are for creating the appropriate applet using the data in stacked format. The user simply needs to specify the column containing each sample and then specify a *Where* expression to determine the values within the column that are associated with each sample. In this case the sample values to be compared are both in the *Score* column and can be differentiated by the value of the *Exam* column (Yellow or White). By default, the sample values in the resulting applet will be labeled simply by the associated column name. To produce a better display, one can specify separate labels for each sample. In this case, the labels have been specified as *Yellow* and *White*, respectively. The inputs for the less common split format are even easier to specify, as only the two columns of data need to be specified in the proper order.

After pressing the *Compute* button, StatCrunch will produce the applet shown in Figure 2. Instructors can produce randomizations by clicking the *1 time*, *5 times* and *1000 times* buttons. These buttons map to the ideas of step, walk and run. By introducing the applet with the *1 time* button, an instructor can show students that the applet is generating a single randomization in the same manner as they did to complete the activity. When the *1 time* button is clicked, the applet produces a pop up window that shows an animation of the shuffling of the labels and the resulting difference between the mean scores for the two groups after the new associations have been made. This difference value is then dropped onto the number line in the graphic within the applet. As was the case in the activity, the value is color-coded based on its value relative to the observed mean difference. Values that are not larger than the observed difference are shown as gray blocks and values that are larger in magnitude than the observed value are shown as red blocks. Each randomization is also tallied in the table above the graphic. This table shows the number/proportion of mean differences from the randomizations that are below the negative of the absolute value of the observed difference, above the absolute value of the observed difference and the total number falling in either of these regions. When using the applet for a one sided test, careful consideration should be given to which table elements are important based on the type of test being conducted (upper tailed, lower tailed or two-sided).

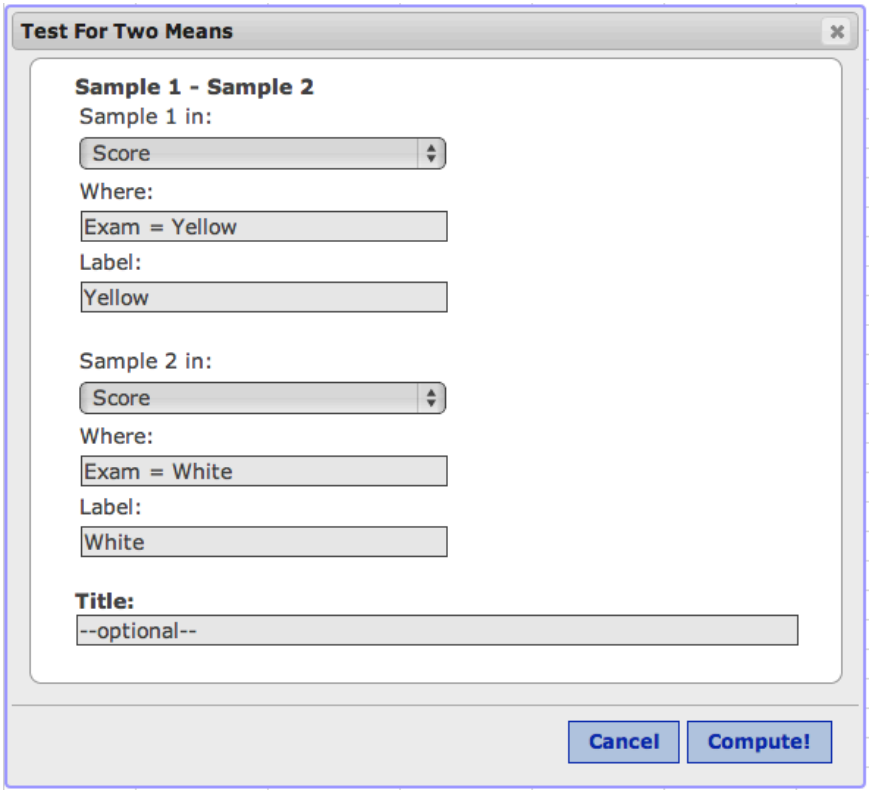

Figure 1: StatCrunch dialog window for randomization test

In this example, the total tally is used since the scenario described is two-sided in nature. The instructor can continue to warm students up to the technology using the *5 times* button which will sequentially add the results to the output from five more randomizations in a less animated fashion. Once students are familiar with the technology, the instructor can then press the *1000 times* button repeatedly to add the results of thousands of randomizations very quickly. The applet contains a listing of all the randomizations that are produced under the *Runs* heading. Any of the randomizations can be investigated by clicking on the corresponding run number. A dialog window will appear showing the corresponding shuffle of the labels along with the means for each sample and their difference. The user can also interact with the graphic by clicking on any of the bars and choosing a run number from the resulting listing of all runs contained in that bar. This method is particularly useful for investigating the randomizations that produce the most extreme differences between the two means.

Once the instructor has generated enough randomizations with the applet, the class discussion can return to the evaluation of the observed mean difference of 6.3. Using the results from the applet, a much more precise estimate of the P-value can be obtained. In the results shown in Figure 2, it appears that there is roughly a 26% chance of observing a difference as or more extreme than 6.3 due purely to random chance. Since these chances are quite high, there is not enough evidence to say that the two different versions of the exam produce significantly different results. In other words, it is entirely possible to see such a difference in means due purely to random chance. After completing this discussion, the instructor may then ask students what sort of percentage might lead them to the opposite conclusion where the two versions of the exam do produce significantly different results. This follow up discussion can be used to reinforce the skeptical approach taken with hypothesis testing where significant differences are those that are very unlikely to occur due to random chance.

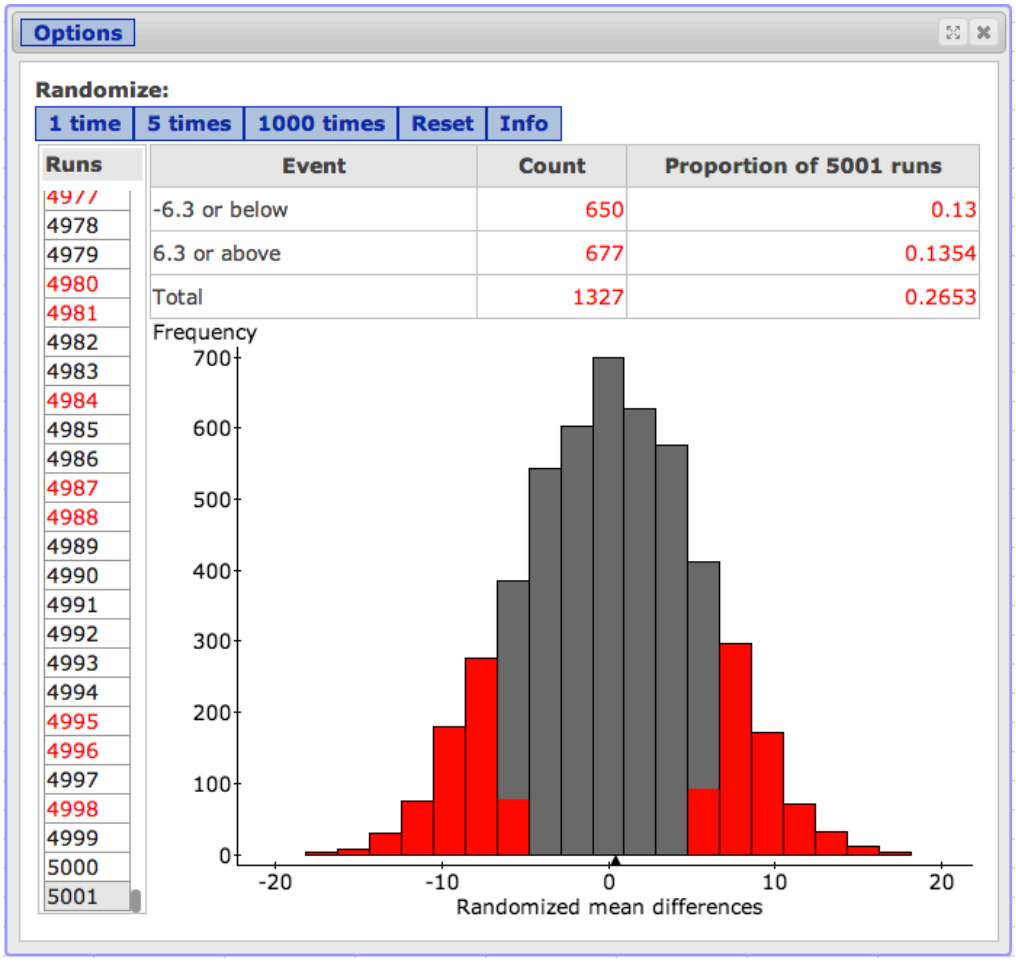

Figure 2: StatCrunch Randomization Applet

DISCUSSION

For the example above, the shuffling of exam labels with scores combined with the resulting tallies clearly promotes the idea that the goal is to compute the likelihood of the observed mean difference (or something more extreme) if there is really no connection between the student's score and their version of the exam. Indeed, these ideas are promoted without any of the standard machinery involving standard errors or standard scores that are developed over the period of several weeks in a traditional course. With the INCIST materials, these ideas are introduced first via a hands-on activity and then followed up with the appropriate technology. One could clearly consider using the technology without the activity, but this may well lead to the student's viewing the technology as a black box in which case the resampling approach is not likely to lead to more conceptual understanding on the part of the student. The necessary role of the activity portion of this approach is being studied extensively in a current research project at North Carolina State University. Results from this study should be available in the summer of 2014.

I have implemented the applet/activity combination discussed above in my own courses over the past several semesters. These materials have also been presented to other instructors at several INCIST workshops. A number of other activity/applet pairings covering other types of hypothesis tests and also confidence intervals have also been developed as part of the INCIST project and class-tested by myself and other workshop instructors. It is impossible to judge the success of this sort of major curricular overhaul over a short period of time. The proper teaching techniques tend to evolve over time as various approaches are implemented and then adjusted based on student feedback. It also typically takes some time for individual instructors to become comfortable with teaching new techniques. Unfortunately, academia is not well structured for instructor's to devote time to such a risky proposition. This is even more problematic in a field such as statistics where most instructors are not classically trained in the area. Resampling methods are not part of the background of most instructors who teach introductory statistics.

I can say, however, based on my own implementation, the early results have been quite mixed. Wile conversations with individual students have suggested a deeper understanding and interest in the field, the data overall have not been so positive. Despite my best efforts, I have not seen a marked change in the learning outcomes for students in my courses since implementing these methods. As was the case before this implementation, about 75% to 80% of my students answer questions correctly about hypothesis tests and confidence intervals based on resampling techniques. While exam performance is only one measure to be considered, I also have not seen any increase in course satisfaction on student evaluations. In fact, there may have been a small decrease since these methods were implemented. In short, resampling techniques have not proven to be a magic bullet for me. Of course, it was never reasonable to consider that they might be. It is naïve to think that any curricular changes will produce huge changes in student learning. Students in large lecture courses such as mine are almost completely focused on grades and not on developing a deep understanding of the subject. Until this culture is changed, the potential gains in student learning/satisfaction from this sort of curricular change will be modest at best. At the end of the day, my students tend to use the technology as a black box in much the same way that they followed the standard recipes to compute results using formulas in years past.

While there has undoubtedly been an uptick in my enthusiasm for teaching the course material using resampling techniques, I have also encountered a number of issues related to the use of these methods. Assessment is certainly more of a challenge when teaching these techniques. The standard exam exercise where summary statistics are provided and students are asked to conduct a hypothesis test is not possible with the resampling approach where complete data is required. This may in fact be the biggest drawback to teaching these methods at the introductory level. Many research papers and online publications still only report summary statistics such as means and standard deviations rather than raw data. Without these standard questions, the instructor is forced to develop new assessment techniques. I opted to adapt this sort of exercise to include raw data. My students then use their own computing equipment to apply the proper resampling techniques to answer these questions in a closely monitored environment. Cobb's statements about the ease of computation in today's technology environment have certainly proven true. My students access the StatCrunch applets on smartphones and tablets as well as laptop computers for both in class activities and exam purposes. This aspect of the implementation has

been remarkably smooth with almost no student issues. In addition to these exercises that require student computation, I also typically include questions based on the StatCrunch applet output in graphical form where students focus simply on interpreting the output correctly.

In terms of other issues, bootstrapping has also proven to be much more difficult to implement than randomization tests. Many instructors from the INCIST workshops have also reported similar issues with introducing the bootstrap in their courses. While randomization provides an interesting intuitive approach for two sample hypothesis tests, the general feeling from both students and instructors has been that the bootstrap does not offer the same type of advantage when it comes to confidence intervals. Treating the sample as a population and resampling from it (with replacement) appears to be less intuitive for most people. The machinery required to develop the bootstrap in the variety of settings encountered in an introductory statistics course is also quite substantial. The hands on activities associated with the bootstrap approach have been difficult to design and conduct in the classroom. The introduction of the bootstrap is required if one wishes to use the resampling approach in the one sample setting where more intuitive randomization techniques do not apply. This should be a major consideration for any instructor who is considering adopting a pure resampling approach for their course. Issues with the bootstrap may be one reason that many instructors have opted for a hybrid approach where randomization is used to motivate hypothesis testing but the standard inferential methods are also covered.

All issues aside, one argument for adopting a resampling approach for introductory statistics is that it should free up a great deal of time in the curriculum since the extensive development of normal distribution theory is not required. This may be true, but after considering this carefully over the last several years, I have arrived at the conclusion that the same sort of time can be saved with a modified standard approach that emphasizes technology for computation. More recently, I have begun to consider the overstated role of inference in introductory statistics. In the data age, students have access to a few data sets with millions of observations and millions of data sets with a few observations. In both cases though, the data sets that are readily available and of interest to students are not appropriate for inferential methods. Data from well-designed experiments and simple random samples are very difficult to come by. Perhaps the reform of the introductory statistics curriculum should be concentrated on helping students work more effectively with the data sets that they may commonly encounter. Such efforts might include a focus on developing better skills for doing basic data manipulations such as filtering and grouping and an emphasis on exploratory data analysis and story telling using visualization techniques.

REFERENCES

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1), 1-15.