USING BOOTSTRAP DYNAMIC VISUALIZATIONS IN TEACHING

Joss Cumming, Christine Miller and Maxine Pfannkuch
Department of Statistics, The University of Auckland, New Zealand
j.cumming@auckland.ac.nz

The increasing availability of technology means that computationally intensive methods such as bootstrapping are now accessible to students. At the university level, where introductory classes are large (about 450 students per class), we introduce students to the idea of a confidence interval using the bootstrap method before they meet the traditional approach later in the course. In this paper we describe our teaching approach using the Visual Inference Tools (VIT) software, which was designed to enhance introductory statistics students' conceptual understanding of bootstrap confidence interval construction. Using data from research we present some of the issues that arose in students' reasoning processes. The implications of this initiative to improve how statistics is taught and to use technology in a way that improves students' understanding will be discussed.

INTRODUCTION

Limited research has been conducted on students' understanding of confidence intervals using the traditional normal-based approach with no apparent research being conducted using the bootstrap method. Considering the ubiquity of confidence intervals within the statistics discipline it is surprising that such a fundamental concept has escaped the attention of researchers. The focus of this brief paper is limited to a succinct description of an introductory approach we use for bootstrap confidence interval construction using specially created dynamic visualizations named *Visual Inference Tools* (VIT) (see: http://www.stat.auckland.ac.nz/~wild/VIT), which are part of the data analysis tool *iNZight*. Using data from a large number of introductory statistics students' post-tests (from n=207 to 710), we explore briefly the following research questions and raise potential issues. From a two-lecture introduction to bootstrap confidence intervals:

- How do students interpret bootstrap confidence intervals?
- What key ideas about the bootstrap process do students mention?

BACKGROUND

Confidence Intervals

In most undergraduate introductory statistics courses the conceptual foundations underpinning confidence intervals are the normal distribution, the Central Limit Theorem and the sampling distribution of estimators. Although confidence intervals have been taught for years in introductory statistics courses, very little research has been conducted on students' understanding of them (Sotos, Vanhoof, Noortgate, & Onghena, 2007). The sparse research that has been conducted on student understanding of confidence intervals has thrown up a raft of misconceptions. These misconceptions include the thinking that a 95% confidence interval (for a mean) contains the plausible values for the *sample* mean, covers 95% of the sample, is the range of individual scores, increases in width with sample size or is not affected by sample size (Fidler, 2006). Garfield, delMas and Chance (1999) listed 13 points for a one-sample situation confidence interval that students should understand, three of which are:

- A confidence interval for a population mean is an interval estimate of an unknown population parameter (the mean), based on a random sample from the population.
- A confidence interval for a population mean is a set of plausible values of the true population mean that could have generated the observed data as a likely outcome.
- The level of confidence tells the probability the method produced an interval that includes the true population mean.

Inherent in these three points are underpinning concepts such as sampling from populations and sampling variability and that making decisions under uncertainty involves conceiving data-generating situations as random processes. Other conceptual underpinnings behind sample-to-population inference involve students understanding that instead of dealing with only one possible *reality*, consideration must be given to a number of possible realities which requires imagining 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 [© 2014 ISI/IASE]

repeated collecting of samples such that the values of the recorded statistic of each sample form a stable distribution of possibilities. That is, understand that a sample can be taken to learn about some characteristics of an unknown population distribution and how calculated statistics from the sample can be used, for example, to produce a confidence interval of plausible values for a population parameter. Students also need to understand why statisticians have confidence in this method for estimating a parameter. We believe that a good understanding of confidence intervals cannot be addressed in a single course of instruction and that students should be exposed to ideas of sampling variability and intuitive confidence interval ideas over several years.

The Bootstrap Method

The current literature in statistics education on the bootstrap is centred on explaining the method, giving teaching examples, and arguing that the method will give students better access to ideas that underpin inference. There is no research to date on the bootstrap method's effectiveness in improving student learning, on students' reasoning using the bootstrap or on any learning issues. Therefore in this section we will discuss the rationale for using the bootstrap method.

In 1979 Brad Efron produced a landmark paper on the bootstrap method that has revolutionized the practice of statistics. Efron's idea was to estimate a sampling distribution from just one sample. By treating this one sample as if it were the population and mimicking the data production process (Hesterberg, 2006), multiple re-samples of the same size as the original sample are taken with replacement from this original sample with the statistic of interest being calculated each time. The variation in estimates from re-samples from this original sample approximate the variation in estimates that would be obtained if many samples from the population were taken. Since the bootstrap method is capable of generating bootstrap distributions for summary statistics such as medians, quartiles, measures of spread, and correlations, it goes far beyond the scope of classical mathematical methods simple enough to be commonly taught. The bootstrap, which has had a major effect on the practice of statistics, is not currently part of the introductory statistics curriculum. With computing power now available to students, "there is no excuse" (Cobb, 2007, p. 13) not to introduce students to the bootstrap method and to correct the mismatch between statistical practice and the introductory curriculum.

Apart from the fact that bootstrapping is rapidly becoming the preferred way to do statistical inference, there are strong pedagogical arguments for introducing the bootstrap into the curriculum. First, the bootstrap can be used to make the abstract concrete by providing "visual alternatives to classical procedures based on a cookbook of formulas" (Hesterberg, 2006, p. 39). These visual alternatives have the potential to make the concepts and processes underpinning bootstrap inference transparent, more accessible, and connected to physical actions. Second, students experience a set of general approaches or a method that applies across a wide variety of situations to tackle problems rather than learning multiple and separate formulas for each situation. Such formulas work in special circumstances but the general approach works in most situations and sometimes it is the only option. Third, for the majority of students, who will never need to study analytic methods, simulation methods such as the bootstrap should be promoted as the only method, as mastery of algebraic representations is not a prerequisite (Wood, 2005).

METHOD AND TASKS

Using Hjalmarson and Lesh's (2008) design research principles, the development process in this project involved two research cycles with four phases: (1) the understanding and defining of the conceptual foundations of inference, (2) development of learning trajectories, new resource materials, and dynamic visualization software, (3) implementation with students, and (4) retrospective analysis followed by modification of teaching materials. The research was conducted over two years and went through two developmental cycles. In the first year, there was a pilot study involving ten students while in the second year, the main study involved 2765 students (fourteen Year 13 classes (last year of high school), seven introductory university classes (about 450 students per class), one workplace class). About 50% of these students may have experienced a normal-based approach to confidence intervals in their last year of high school. None of them had previous experience of the bootstrap method. Some of the data collected were pre- and post-tests of all the students and pre- and post-interviews with 38 students. There were two versions of the post-test in the main study, a bootstrap post-test and a randomization post-test (not reported here) to which students were randomly allocated. For the tests, assessment frameworks were developed from the student data for 11 free-response questions. Two researchers coded about 100 free-response questions and came to a consensus and then two research assistants coded the data.

Since our introductory statistics courses at the university level cater for a large number of students each year (about 5000) with many competing needs from client departments, it was only possible to collect data on a short learning trajectory, which briefly introduced students to the bootstrap method in two lectures at the beginning of the course. Students returned to the method and ideas later in the course when the traditional approach was introduced. School students were similarly restricted because the bootstrap method was not part of the school curriculum that year.

The introduction of the bootstrap method necessitated the development of hands-on activities, dynamic visualizations for learning and analysis (*VIT*), and the creation of new verbalizations. We will now briefly describe two modules of the *VIT* package (Figures 1 & 2).

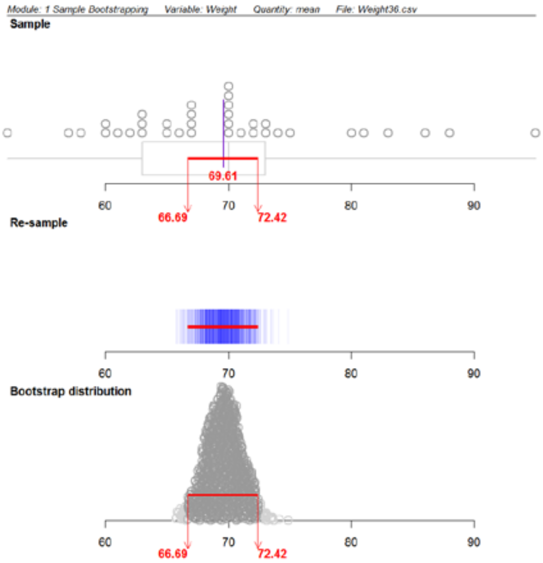

Figure 1. *VIT* bootstrap confidence interval module

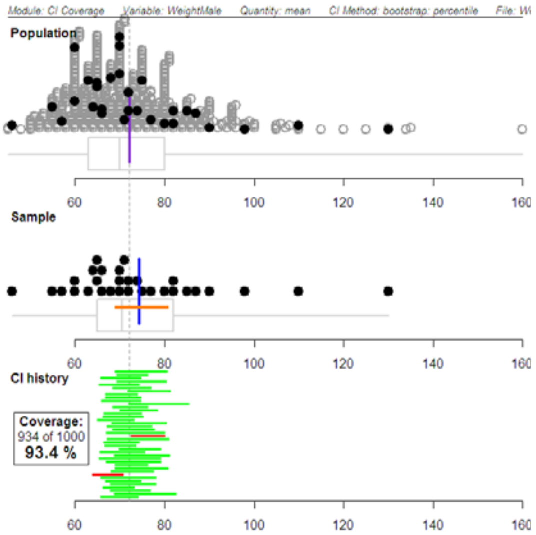

Figure 2. *VIT* bootstrap confidence interval coverage module

The bootstrap confidence interval construction module graphics window comprises three plots: Sample, Re-sample and Bootstrap distribution (Figure 1). The Sample plot displays the original sample, in this case the weights of a randomly selected sample of university students. One of the observations is randomly selected and highlighted in the Sample plot and then a copy of this observation drops into the Re-sample plot. This is repeated until the re-sample is of the required size, generating a temporary dot plot. The re-sample mean is calculated and shown on the re-sample dot plot as a permanent vertical blue line. As each re-sample mean is displayed, a copy of it drops down onto the Bootstrap distribution plot. This process is repeated 1000 times building up a band of blue re-sample means in the Re-sample plot and a dot plot of re-sample means in the Bootstrap distribution plot. Finally, a percentile bootstrap confidence interval, using the central 95% of the bootstrap distribution, is displayed and superimposed on all three plots. The coverage module (Figure 2) involves using a population such as weights of university students, taking 1000 re-samples from a sample of size 30, for example, to create a bootstrap confidence interval for the mean and checking whether the interval has captured the population mean. This procedure is then repeated 1000 times and the percentage of times the true population mean is captured is recorded.

The two-lesson learning trajectory is situated in contexts that motivate the need for finding an uncertainty band around an estimate and the “big ideas” behind the bootstrap method. It consists of the following: a hands-on activity of weights of university students with a sample of size nine that progressed the students through each stage of the bootstrap method using the same representations, as shown in Figure 1, through to a *VIT* demonstration using 1000 re-sample medians; bootstrap confidence intervals for comparison of medians and means; and the concept of coverage of the interval. Three weeks later, after the students had handed in an assignment requiring the use of the *VIT* confidence interval module, they sat the unannounced post-test.

RESULTS

In this paper we report results from three questions from the bootstrap post-test and reference one follow up interview to illustrate a particular issue that arose from one question. Note that currently not all of the data has been coded so the number of responses varies for each question and students who did not respond to the question have been excluded.

Research Question 1: How Do Students Interpret Bootstrap Confidence Intervals?

For the first post-test question, which was part of a longer question, students were asked to interpret the bootstrap confidence interval in Figure 3, which was derived from the weekly incomes of a sample of 21 New Zealanders who worked full-time and had a bachelor's degree. Of those who responded (n=710), 2.0% gave idiosyncratic responses, 20.3% read the data rather than interpreting it (e.g., *"the bootstrap confidence interval is from 693.3 to 937.1"*), 21.6% interpreted the data in terms of weekly income rather than *mean* weekly income, while 56.1% interpreted the interval correctly. For those 77.7% of students who interpreted the data, 80.4% used language such as *"it is a fairly safe bet"*, which was used in class to indicate specifically the notion of coverage but, for the students the phrase may simply indicate a recognition of the uncertainty present when drawing a conclusion using a confidence interval or may relate to the fact that the middle 95% of the bootstrap distribution means/medians were taken to form the confidence interval.

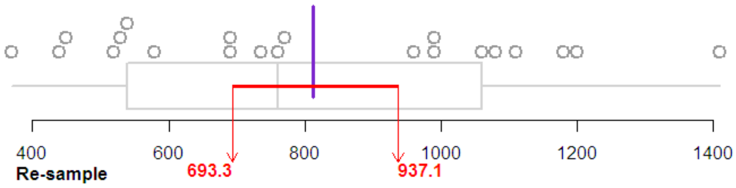

Figure 3. Bootstrap confidence interval plot

An illustration of why students may not be using the word *mean* in the interpretation of the interval occurred in the following student interview. For a prior item, the student was simply asked to explain how he would label the bootstrap distribution *x*-axis (see Figure 4). On responding that the label would be the *mean* weekly incomes, he immediately recognized that his answers to several questions were wrong and he changed them, including his confidence interval interpretation, without further prompting. It seemed that the interview prompt had made him realize that the use of the word *mean* was critical for interpretation.

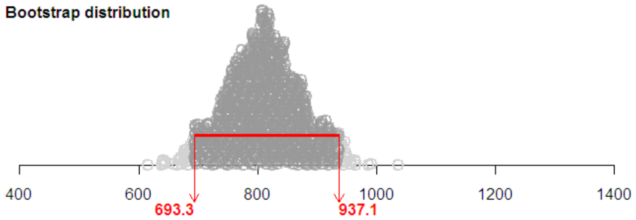

Figure 4. Bootstrap distribution

After students were asked to interpret the bootstrap confidence interval in Figure 3, they were given the following second post-test question: "Does Ross's confidence interval contain the actual mean weekly income of New Zealanders who worked full-time and had a bachelor's degree? Explain your reasoning." In class students were given a similar multi-choice clicker question except it used the word *population* rather than the word *actual*, and the choices were Yes/No/Don't Know. The expectation was that the students would respond with Don't Know and with the coverage reason that was given in class. Of those who responded (n=268), 7% gave an idiosyncratic or incorrect response, 14% said the interval did not contain the actual mean, 24% said it did, while 55% responded with Don't Know.

For those 14% who responded with No, concerns were about the fact the data were: only from a sample *("No. It only shows the mean of a select sample of the population")*; not representative *("No, because this is an estimate and cannot be representative of the population of New Zealanders")*; based on a small sample size *("No. A sample of 21 New Zealanders is such a small sample to be making any assumptions from")*; and not the population *("No. We cannot obtain the actual mean without studying the whole population")*. The proportion of 24% who said the interval contained the actual mean comprised 2% of the students, who did not seem to consider the possibility that the interval may not contain the actual mean *("Yes, because the confidence interval gives us a plausible range of figures")*, and 22%, some of whom may have misinterpreted the question and thought they were being asked to state whether the sample mean was in the interval *("Yes, the mean of $812 is within the confidence interval of $693.3 and $937.1")*. The remaining 55% responded with Don't Know and comprised 47% of the students who invoked chance ideas *("We don't know. We can only say that it is highly likely that the actual mean falls within this confidence interval")*, 5% who expressed uncertainty and the need to take data from the whole population to find the actual mean *("We do not know. Unless we took a census we have no way of knowing what the actual mean weekly income is, therefore all estimates are uncertain")*, and 3% who expressed some coverage ideas that the bootstrap method produces a range of values that contain the true unknown value of the population mean most of the time that we use it *("[bootstrap] confidence intervals encompass the mean/median 96.8% of the time"* and *"We do not know, but there is a large chance that the interval does contain the actual mean weekly income because when properly done it [the bootstrap method] works approximately 96% of the time")*.

From these and other findings (not reported) two issues arise: (1) student understanding of the requirements of the question; that is, understanding the use of the language "interpret" and "actual mean"; and (2) students' understanding of the bootstrap confidence interval and the language used to convey the concepts. Within this second issue there seem to be four potential considerations about students' understanding of bootstrap confidence intervals. The students:

- Do not know that the confidence interval is giving an estimate for the population mean and think it gives a range of possible sample means or individual values (cf. Fidler, 2006).
- Do know that the confidence interval is giving an estimate for the population mean but fail to appreciate that the lack of the critical word *mean* changes the interpretation.
- Believe that a random sample cannot tell them about some characteristics of the unknown population based on reasons such as representativeness or sample size, that is, the concepts behind sample-to-population inference have eluded them.
- Have not cognitively integrated the multiple ideas of uncertainty including its quantification.

With respect to uncertainty students need to coordinate multiple perspectives of samples such as variability of individuals in a sample, of samples themselves, and of the calculated statistics from sample to sample. For the quantification of uncertainty with the bootstrap confidence interval students need to coordinate the information from the uncertainty band constructed around the estimate using the middle 95% of bootstrap distribution values and empirical coverage information about the quantification of the uncertainty for the method itself. The ability to cognitively link the *VIT* modules for bootstrap confidence interval construction and coverage and what uncertainty the language "it is a fairly safe bet" refers to amongst all the types of uncertainty present within the bootstrap process will take extended time and learning. A hands-on activity for the coverage idea would be an appropriate learning experience to develop with specific links to language.

Research Question 2: What Key Ideas About the Bootstrap Process Do Students Mention?

For the third post-test question students were asked to explain one key idea underpinning the bootstrap process to estimate a parameter. About 32.9% of the students who responded (n=207) could verbalize the "big ideas" of a connection between multiple re-sampling from a population and multiple re-sampling from a sample, with a small percentage of them (2.5%) specifically mentioning that the variability in the re-sample means mimics the variability in the means from multiple population samples. The rest of the students focused on describing the bootstrap procedure (50.7%) and/or mentioned other key ideas such as all estimates are uncertain, the bootstrap method works most of the time, or that samples must be random. While it is pleasing to see that a number of students are beginning to think beyond the mechanics of obtaining a bootstrap confidence interval, there is an issue about how to draw students' attention and orientation towards grasping the big ideas. More dynamic visual imagery can be designed to link multiple sampling from a population to multiple re-sampling from a sample but the visual argument for similarity in the uncertainty bands obscures a much deeper conceptual inversion argument about the bootstrap confidence interval generated; an issue that has not yet been addressed (see Pfannkuch, Wild, & Parsonage, 2012 for a much fuller discussion).

CONCLUSION

Within the limitations of a two-lecture introduction to the bootstrap method, we believe that the students were beginning to gain an appreciation of some key ideas underpinning sample-to-population inference for the construction of a bootstrap confidence interval. Since many students referred to notions such as multiple re-sampling, that all estimates are uncertain, and could describe the bootstrap procedure, we conjecture that the dynamically linked visualization *VIT* software assisted their thinking and understanding. This *VIT* software makes the underlying processes visually transparent and accessible in a manner that Wood (2005) states simulations should facilitate. Dynamic visualizations, however, cannot assist in the interpretation of confidence intervals with many previously identified misconceptions surfacing (Fidler, 2006). It is within this area that more learning experiences need to be developed and researched. Specifically attention needs to be focused on overcoming some students' reluctance to countenance inference about a population from a sample, which includes the relationship between sample size and confidence interval width. The language of a question and the precise wording for confidence interval interpretation needs to be addressed. Finally, our findings suggest the necessity of untangling the multiple ideas of uncertainty and the quantification of uncertainty associated with confidence intervals, and the linking of the language of uncertainty to specific conceptual constructs.

ACKNOWLEDGEMENT

Work in this paper is supported in part by a grant from the Teaching and Learning Research Initiative (http://www.tlri.org.nz/sites/default/files/projects/9295_summary%20report.pdf).

REFERENCES

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1), 1-15. http://escholarship.org/uc/item/6hb3k0nz

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1-26.

Fidler, F. (2006). Should psychology abandon *p* values and teach CIs instead? Evidence-based reforms in statistics education. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php?show=17

Garfield, J., delMas, R., & Chance, B. (1999). *Tools for teaching and assessing statistical inference*. Retrieved from www.tc.umn.edu/~delma001/stat_tools/

Hesterberg, T. (2006). Bootstrapping students' understanding of statistical concepts. In G. Burrill (Ed.), *Thinking and reasoning with data and chance. Sixty-eighth Yearbook of the National Council of Teachers of Mathematics* (pp. 391-416). Reston, VA: NCTM.

Hjalmarson, M., & Lesh, R. (2008). Engineering and design research: Intersections for education research and design. In A. Kelly, R. Lesh, & K. Baek (Eds.), *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching* (pp. 96-110). New York: Routledge.

Pfannkuch, M., Wild, C. J., & Parsonage, R. (2012). A conceptual pathway to confidence intervals. *ZDM – International Journal on Mathematics Education*, *44*(7), 899–911. doi: 10.1007/s11858-012-0446-6.

Sotos, A., Vanhoof, S., Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review 2*, 98–113.

Wood, M. (2005). The role of simulation approaches in statistics. *Journal of Statistics Education*, *13*(3), 1-11. www.amstat.org/publications/jse/v13n3/wood.html