Developing Statistical Inferential Concepts in Introductory Courses

Stephanie Budgett and Maxine Pfannkuch
University of Auckland, Auckland, New Zealand
Corresponding author: Stephanie Budgett, e-mail: s.budgett@auckland.ac.nz

Abstract

Hypothesis testing and confidence intervals are recognized as difficult areas for students of introductory statistics, partly due to the reliance on abstract mathematical concepts in the traditional approach to teaching statistical inference. Given the recent advances in computing power, we are now able to harness new technologies and make use of computer intensive methods and use visual rather than mathematical approaches to develop students' understanding of statistical inference. Rising to George Cobb's challenge to place the logic of inference at the heart of the introductory statistics curriculum, a large collaborative project explored new ways of introducing final year secondary school and first year university students to inferential reasoning. The project involved using innovative dynamic visualizations using for the bootstrap and randomization methods to teach statistical inference. Of interest was to establish whether this new approach, using hands-on activities and visualizations, facilitated students' conceptual access to the logic of inference. In this paper we provide a brief overview of the dynamic visualization software that was developed for the randomization method. Using responses from students to two written randomization method assessment items we discuss some issues that have arisen as a result of teaching statistical inference in new and innovative ways.

Keywords: Uncertainty, randomization method, dynamic visualizations

1. Introduction

Traditionally, methods of teaching statistical inference in introductory courses have been grounded in mathematical theory and formal calculations. Such an approach, with its reliance on abstract mathematical concepts, has placed obstacles in the path of student understanding (Chance, delMas, & Garfield, 2004). It is well-known that most students have trouble following the unfamiliar logic associated with statistical reasoning taught in this way, and that the majority of students emerging from traditional introductory courses in statistics fail to have a sound understanding of statistical inference (Nickerson, 2004; Rossman, 2008).

Changes are afoot, however. As George Cobb noted, "…computers are changing the teaching of our subject" (Cobb, 2007, p. 1). Re-sampling techniques such as the bootstrap and randomization methods, although algorithmically simple, are more powerful than normal-based theoretical methods, are unencumbered by distributional (and other) assumptions and are applicable to many different situations. Furthermore, these methods show promise as tools for facilitating better conceptual understanding of statistical inference in introductory statistics students, largely attributable to the fact that they lend themselves to visual processes which may help to consolidate some of the core underpinning concepts of statistical inference (Budgett, Pfannkuch, Regan, & Wild, in press).

Re-sampling techniques are commonly used in statistical practice and are now permeating the field of statistics education, with several groups underpinning their introductory courses with the randomization method (Gould, Davis, Patel, & Esfandiari, 2010; Rossman, 2008; Tintle, VandenStoep, Holmes, Quisenberry, & Swanson, 2011). Our research project had the aim of discovering if and how dynamic visualizations using the bootstrap and randomization methods can allow introductory

students more access to the big ideas of statistical inference. This paper will deal with some issues that have arisen as a result of the introduction of the randomization method.

2. Background to the project

A large team contributed to the development of new and innovative approaches to teaching statistical inference using a design research approach. A fuller description is given in Budgett et al (in press). Briefly, the conceptual foundations of inference were defined in conjunction with new resource materials incorporating teaching sequences, dynamic visualizations and assessment items. Modification and supplementation of resource materials was carried out in light of feedback from a small pilot study of ten students. Thereafter the new approach was implemented with over 2,700 final school year and university introductory statistics students.

3. Description of teaching sequence

Students were introduced to the concept of the randomization method following a series of lectures covering experimental and observational studies. In addition, instructors provided stories which described the process of statistical inferential argumentation that we might carry out on a routine basis in our day-to-day lives (Vickers, 2010). Students were exposed to the concept of chance acting alone using a software module which was developed to demonstrate the types of differences that one might experience by simply randomly allocating a set of observations to two groups without actually administering any form of treatment. The rationale for using this module was that in the pilot study it became clear that students were confused by the concept of *chance acting alone*. Figure 1 illustrates the weights of 30 people and their subsequent random re-assignment to two groups. Differences in the mean weights of the two groups are recorded in the middle panel, with these differences then dropping down to build up a re-randomization distribution in the bottom panel. Students could then see that the absolute differences in the mean weights between the two groups can be up to 10kg, simply by *chance acting alone*. Hence we hoped that this module would prompt students to consider chance explanations when observing differences between two groups.

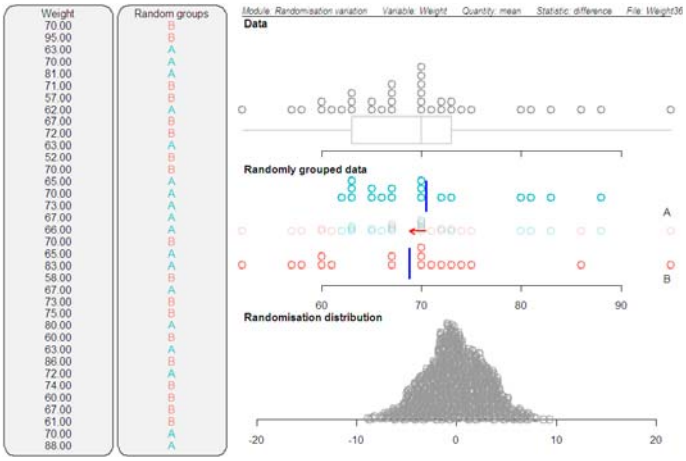

Figure 1. Random re-allocation under chance acting alone

The randomization method was then introduced via the description of a randomized experiment designed to investigate if a series of special exercises administered by caregivers would have any impact on the age at which a baby started walking (Zelazo, Zelazo, & Kolb, 1972). The walking ages of two groups of babies were recorded, one group having been randomly allocated special exercises, the other group randomly allocated as controls. It was noted that the difference in the median walking ages between the two groups (control minus exercise) was 2.25 months.

A hands-on activity followed, enabling students to actively re-randomize the

babies, using cardboard tickets, to the two groups and to record the difference in average walking ages in order to experience what differences between group averages might be typical through chance acting alone. These re-randomized differences were collected and plotted by the instructor. Students repeated this process several more times, at which point visual inference tools (VIT) (http://www.stat.auckland.ac.nz/~wild/iNZight/dlw.html) were introduced to automate the process. The software was designed to closely mimic the hands-on activity, with a vertical arrangement of the graphics panel within the dynamic visualization tool (see Figure 2). For example, the top panel of Figure 2 represents the observed data with a difference in mean walking ages between the control and exercise groups of 2.5 months, the middle panel illustrates one re-randomization which has produced a smaller difference in means but in the opposite direction, while the third panel shows the distribution of 1000 re-randomizations. We conjecture that the fact that the entire process can be viewed within the same screen may lead to a more concrete understanding of what the re-randomization distribution represents.

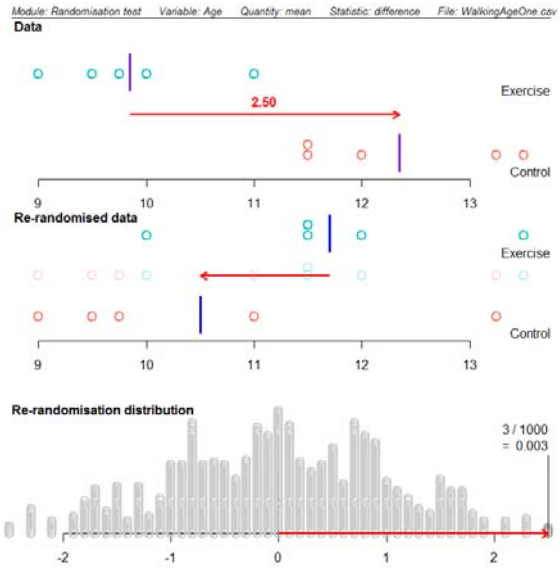

Figure 2. Vertical arrangement of dynamic visualization for re-randomization

Once a distribution of re-randomization differences has been built up, it is possible to visualize the observed difference between the two groups, i.e. 2.5 months, in relation to the re-randomization distribution. The fact that the observed difference of 2.5 months lies way out in the upper tail suggests that such a difference, although possible, is highly unlikely if chance was acting alone. In fact, under chance alone, an observed difference of at least 2.5 months occurred only 3 times out of 1,000 re-randomizations, giving rise to a tail proportion of 3/1000 = 0.003. Thus we would make the claim that another factor is acting alongside chance (i.e. the special exercise programme) to explain the observed difference. If the observed difference did not lie out in the tails of the re-randomization distribution, then our conclusion would be that we have no evidence to discredit the chance explanation and that there may well be a treatment effect but that are unable to detect it from the obscuring effects of chance variation. Thus the call we would make would be that chance *may* be acting alone, or some other factor *may* be acting alongside chance, we cannot tell either way.

The second randomized experiment introduced to students, comparing walking ages of the special exercise group and another group, resulted in an observed difference of 1.4 months corresponding to a tail proportion of about 15%. Being mindful of a common misconception to regard a chance explanation as the only explanation, discussion centred around the fact that chance *could* be acting alone, or the special exercise programme *could* be effective, and that we had insufficient evidence to say either way. We anticipated that tying such a conclusion back to the

statistical argumentation used in everyday life might allay the misconception that a large tail proportion (or *p*-value) gives evidence in support of the chance explanation. Students were provided with a guideline stating that a tail proportion of less than about 10% suggested evidence of a treatment effect.

4. Results

Prior to the introduction of the randomization method, a pre-test was administered to over 2,700 students. One of the pre-test questions described the results of a randomized experiment in which researchers randomly assigned 14 male volunteers with high blood pressure to one of two four-week diets: a fish oil diet and a regular oil diet (Knapp & FitzGerald, 1989). Each participant's blood pressure was measured at the beginning and end of the study, and the reduction in blood pressure was recorded. Plots of the data demonstrated that the reduction in blood pressure values for the fish oil group tended to be greater than those for the regular oil diet (see Figure 3).

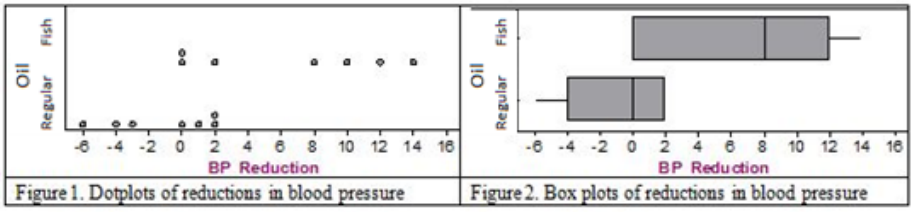

Figure 3. Part of one pre- and post-test question

Two possible explanations for the observed difference are that (1) the treatment is effective and fish oil results in a reduction in blood pressure and (2) that chance is acting alone and it just so happened that those who experienced greater reductions in blood pressure ended up in the fish oil group. When asked for two possible explanations for the observed difference, many students (75.4% of the 1886 who responded to this question) were able to provide a response that demonstrated that they were considering treatment ideas at some level (see codes ***T*** and ***MT*** in Table 1).

Table 1: Coding framework for one pre-test question

Code	Treatment Explanation		Chance Explanation	
	Code	N	Code	N
Treatment or chance ideas present	T	939	C	84
Moving towards treatment/chance ideas	MT	483	MC	483
No treatment or chance ideas present	NT	467	NC	911
No response	NR	879	NR	1287
Totals		2765		2765

Two-thirds of the students who provided some level of treatment explanation were able to state that the observed difference was attributable to the fish oil treatment, coded as ***T***, with responses such as:

S1 Being on a fish oil diet reduces blood pressure in study volunteers. ***T***

Students whose responses fell into the ***MT*** category simply made a correct and relevant observation regarding the difference between the treatments, but fell short of attributing the difference to the treatment, with responses such as:

S2 The median value is greater for fish oil than for regular oil. ***MT***

Such responses were coded ***MT*** since we believed that these students had the idea that the treatment was somehow connected to the observed results and they were therefore moving towards the concept of associating the observed difference with treatment.

Fewer respondents (38.3%) were able to demonstrate a consideration of chance ideas (***C*** and ***MC***), with most considered moving towards chance with responses such

as:

S3 The time of day the blood pressure was taken. ***MC***

We considered such a response, indicating that the student is considering other variables that might explain the observed between the two groups, as moving towards chance ideas since any imbalances between the treatment groups must have arisen by chance owing to the study design. Whether this is what the student intended is another matter. Thus from the pre-test assessment, it would appear that chance ideas, whether directly or indirectly, are not being considered by the majority of students.

One week after the teaching sequence, a post-test was administered to all students. Note that owing to time constraints two versions of the post-test were developed: a bootstrapping post-test and a randomization post-test, to which students were randomly allocated. There were some common pre-test items in both versions of the post-test. The randomization post-test was completed by half of the students, with 816 having matched pre- and post-tests. An analysis of the post-test responses to some of the questions demonstrated that more students are now considering chance-related explanations (74.3% of those who responded), with many of these students considering chance explanations alongside a treatment explanation. More students than in the pre-test (85.7% of those who responded) provided a treatment explanation at some level, with 80% of those coded as ***T*** responses.

Table 2: Comparison between pre- and post-test treatment and chance explanations

		Post-test Treatment							**Post-test Chance**				
		T	*MT*	*NT*	*NR*	*Totals*			*C*	*MC*	*NC*	*NR*	*Totals*
Pre-test Treatment	*T*	134	51	39	62	286	**Pre-test Chance**	*C*	13	6	6	4	29
	MT	84	21	17	20	142		*MC*	59	18	30	33	140
	NT	94	19	6	14	133		*NC*	108	26	56	72	262
	NR	145	28	34	48	255		*NR*	187	46	68	84	385
	Totals	457	119	96	144	816		*Totals*	367	96	160	193	816

We were also interested in students' interpretation of a large tail proportion, particularly in light of the misconceptions associated with interpreting large *p*-values in the traditional normal-based approach to statistical inference (Cohen, 1994; Falk & Greenbaum, 1995). Thus a further post-test question asked students to consider a large tail proportion of 0.3 and to state the conclusion that the researchers should make.

Unfortunately, about half of the students who answered this question misinterpreted the tail proportion of 0.3, with most of these having trouble in converting 0.3 to a percentage.

S4 If the tail proportion was 0.3/ 3% it would mean that chance isn't acting alone. This means they have evidence against chance acting alone

Several issues were identified for those students who interpreted the tail proportion correctly numerically. A small number of these students viewed the tail proportion as the probability that chance is acting alone, with comments such as:

S5 There is a high chance (30%) that chance is acting alone.

Thus it would appear that misconceptions associated with the interpretation of large *p*-values are still prevalent which is not surprising given the nature of the argumentation has not changed.

Of students who responded without incorrectly stating or implying that 0.3 was bigger than 10%, only 15% were able to articulate a high level interpretation of a large tail proportion. It would appear that there remains confusion around the concept of chance acting alone and the concept of chance acting alongside treatment. Some of the better responses expressed the idea that there may be a chance component, or a treatment component, and that we are unable to conclude which.

S6 Fish oil may cause a change in blood pressure, however there is not sufficient evidence to prove that it is not caused by chance.

However, from such a response it is not clear if there is awareness of the fact that chance is always acting and that there is insufficient evidence to state that the difference is not due to the fish oil treatment *plus* chance.

6. Conclusions

The traditional approach to teaching statistical inference, grounded in mathematical theory and probabilistic reasoning, has resulted in difficulties for the majority of students of introductory statistics. The randomization method and associated dynamic visualizations have the potential to clarify many of the underpinning concepts of statistical inference, in particular the behavior of the chance alone phenomenon. We believe that the randomization approach, accompanied by dynamic visualizations which enable students to experience the chance alone phenomenon, facilitate understanding of the concept that chance is always acting, and that this needs to be considered when looking for evidence of a treatment effect.

However, the nature of the argument remains a problem for many students. In the same way that misconceptions arise from the interpretation of a large *p*-value within the traditional framework, students exposed to the randomization method have trouble interpreting a large tail proportion. While the visualizations have clarified the way in which a tail proportion is obtained, and relate the tail proportion to an understandable distribution, the indirect nature of the interpretation still remains. Thus further efforts are required to develop students' reasoning processes when arguing under uncertainty.

References

Budgett, S., Pfannkuch, M., Regan, M., & Wild, C. J. (2013). Dynamic Visualizations and the Randomization Test. *Technology Innovations in Statistics Education* (in press).

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi, & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Cohen, J. (1994). The earth is round (p<0.05). *American Psychologist, 49* (12), 997-1003.

Falk, R., & Greenbaum, C. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology , 5* (1), 75-98.

Gould, R., David, G., Patel, R., & Esfandiari, M. (2010). Enhancing conceptual understanding with data driven labs. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute.

Knapp, H., & FitzGerald, G. (1989). The antihypertensive effects of fish oil. A controlled study of polyunsaturated fatty acid supplements in essential hypertension. *New England Journal of Medicine , 321* (23), 1610-1611.

Nickerson, R. (2004). *Cognition and Chance.* Mahwah, NJ: Lawrence Erlbaum Associates.

Rossman, A. (2008). Reasoning about informal inference: One statistician's view. *Statistics Education Research Journal , 7* (2), 5-19.

Tintle, N., VandenStoep, J., Holmes, V., Quisenberry, B., & Swanson, H. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education , 19* (1).

Vickers, A. (2010). *What is a p-value anyway?* Boston, MA: Pearson Education Inc.

Zelazo, P. R., Zelazo, N. A., & Kolb, S. (1972). 'Walking' in the Newborn. *Science, 176*, 314-315.