Building intuitions about statistical inference based on resampling

Jane Watson
University of Tasmania
<jane.watson@utas.edu.au>
Beth Chance
California Polytechnic State University
<bchance@calpoly.edu>

Formal inference, which makes theoretical assumptions about distributions and applies hypothesis testing procedures with null and alternative hypotheses, is notoriously difficult for tertiary students to master. The debate about whether this content should appear in Years 11 and 12 of the *Australian Curriculum: Mathematics* has gone on for several years. If formal inference is not included in Years 11 and 12, what statistical content, if any, should there be? Should students continue learning more data handling skills, which are a feature of the F–10 curriculum (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2011)? Perhaps the focus should be on procedural aspects, such as correlation and lines of best fit, employing principles from calculus. Or perhaps the curriculum should drop statistics and focus on the more complex theoretical aspects of probability.

To imagine a school curriculum without an acknowledgement of some aspects of inference, however, should be impossible for the developers of the curriculum. Statistics is about carrying out investigations with data from random samples to answer questions about populations or with data from randomised experiments to draw causal inferences. At some point in the development of understanding of the inferential process, probabilistic reasoning becomes involved because the decisions about the questions cannot be made with certainty. Although the word inference is not used in the F–10 mathematics curriculum (ACARA, 2011), there are places where the language implies that decisions are in the offing. Terms such as "make predictions" (Year 2), "interpret data sets" (Years 5–7), "investigate" (Years 8–10), and "evaluate" (Year 10) are associated with decision making, as seen in the elaborations in the curriculum. The question that arises in the classroom is how to link students' experiences across the years to the aim of appreciating the inferential nature of statistics in making generalisations based on data. The ingredients are there up to Year 10 but the connections are missing. Should the connections begin to be made in Years 11 and 12?

The statistics education research community has been discussing the lead in to inference, through informal inference, for some time, and for example the fifth Statistical Reasoning, Thinking and Literacy Forum (SRTL-5 in 2005)

had informal inferential reasoning as its theme. The SRTL-6 and a special issue of the journal *Mathematical Thinking and Learning* have been devoted to "the role of context in developing reasoning about informal statistical inference" (Makar & Ben-Zvi, 2011). Over the last few years various phrases have been used to describe the concept. Ben-Zvi (2006) used IIR (informal inferential reasoning), whereas Pratt and Ainley used ISI (informal statistical inference). Makar and Rubin (2009) presented a useful framework that can be summarised as context and a question, where *evidence* is used to make a *generalisation* beyond the data with an acknowledgement of *uncertainty*.

In formal statistical inference, the evidence comes from representative data collected randomly in a sufficient quantity from a population or process satisfying the mathematical assumptions of the statistical model that is to be tested, in order for a specific probability to be placed on the likelihood that the observed data could have arisen from the sampling distribution derived from the null hypothesis. Those who only accept the formal way of considering inference reject informal inference as inadequate for answering meaningful statistical questions, whereas those who adhere to informal inference believe that as long as the evidence and the type of generalisation are made clear along with acknowledging the uncertainty associated with the conclusion, then others are able to believe (or otherwise) the conclusion (Cobb, 2007).

Many examples of students conducting and analysing statistical investigations of an informal nature exist in the literature. From Ben-Zvi's early work with young children (Ben-Zvi, 2006; Ben-Zvi, Gil, & Apel, 2007), through Watson's (2008) work with Year 7 students, to Pfannkuch's (2011) study with Year 10 students, many student experiences have led researchers to believe that important intuitions are being built about the inferential process that will form a solid foundation for formal inference if met in later years of education. If formal inference is not encountered later, at least students should have a grounding that will allow them to ask questions about claims they see/hear in wider society.

The types of investigations that students at school find interesting are often those that involve differences between two groups, for example boys and girls or year levels within the school (e.g., Watson & Wright, 2008). The question of generalising differences observed in the classroom or school can be assisted for example by collecting random samples of data from the Census@School website (www.abs.gov.au). The concept of determining how much difference in the samples can be considered as a genuine difference in the populations evolves over time as students gain experience with analysing graphs of distributions. Initially students are often overly conservative, not wanting to declare a difference genuine unless there is no overlap of two distributions, or the opposite, declaring any slight difference as meaningful (Watson, 2008).

Moving to the stage of informal inference where randomness is deliberately introduced into the study design through random sampling and/or random assignment, *resampling* methods offer a way of evaluating evidence to support a generalisation that can be reported with an associated frequency-based probability. Simon (1997), one of the strong advocates of resampling states, "Resampling refers to the use of the observed data or of a data generating mechanism (such as a die) to produce new hypothetical samples, the results

of which can then be analysed" (p. 2). In particular the available data on a variable can be randomly reallocated either with or without replacement.[1] The process then recalculates the value of the statistic of interest for the resampled data, perhaps the difference in the group means or medians. This is done many times to estimate the relative frequency (probability) with which the original statistic (or more extreme) would be expected to occur compared to the differences with random reallocation. Simon, as well as others, suggest a diversity of data analysis situations where the process can be used to support generalisations from samples. Examples that use resampling without replacement to consider hypothesised differences in two groups are based on comparing proportions from two-way tables (e.g., Rossman, 2008; Scheaffer & Tabor, 2008; Stephenson, Froelich, & Duckworth, 2010) or comparing means or medians for two groups (e.g., Arnholt, 2007; Christie, 2004; Clements, Erickson, & Finzer, 2007; Ricketts & Berry, 1994; Scheaffer & Tabor, 2008; Shaughnessy, Chance, & Kranendonk, 2009; Taffe & Garnham, 1996). Stephenson et al. also compare their results with and without replacement. Examples based on resampling with replacement to mimic the sampling from a population to estimate a statistic include estimation of a mean (e.g., Arnholt, 2007; Christie, 2004), a correlation coefficient (e.g., Arnholt, 2007; Christie, 2004), and a confidence interval (e.g., Engel, 2004; Johnson, 2001; Wood, 2004).

In being exposed to this process, students are initially given hands-on experiences in carrying out the resampling, which is then transferred to a software package or applet to be carried out a large number of times, perhaps 100 or 1000. This process has been used as an adjunct to traditional teaching of hypothesis testing in university statistics units (e.g., Park, delMas, Zieffler, & Garfield, 2011; Reaburn, 2011; Rossman & Chance, 2008; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011). It has also been a feature of Shaughnessy, Chance, and Kranendonk's (2009) *Focus in High School Mathematics Reasoning and Sense Making: Statistics and Probability*. Various software packages can be adapted to handle the resampling process. Taffe and Garnham (1996), for example, provided the instructions to use *Minitab* and Christie (2004) did so for *Excel*. Ricketts and Berry (1994) used the purpose-built software *Resampling Stats* and Arnholt (2007) provided instructions for using R (R Development Core Team, 2004). An investigation titled "Orbital Express," which involves students in dropping two types of paper objects onto a target (Clements, Erickson, & Finzer, 2007), illustrates the process using *Fathom Dynamic Data* (Key Curriculum Press, 2005). Rossman and Chance provide applets for various types of resampling procedures at <www.rossmanchance.com/applets>.

The more recent statistical software from Key Curriculum Press, *TinkerPlots* (Konold & Miller, 2005, 2011), provides an exceptionally user-friendly way of

1 There is some debate among statisticians about the conditions for using resampling with and without replacement and the exact terminology to use. See Stephenson et al. (2010) for a useful discussion and comparison of these two resampling possibilities with comparing two proportions. The resampling method used in the present paper (without replacement) is perhaps more appropriately called re-randomisation or reallocation, because it is carried out without replacement and mimics the randomness inherent in the random assignment process rather than the sampling process from a larger population. The term resampling is used here for simplicity and applies to both resampling with and without replacement.

carrying out resampling and re-randomisation. It creates a constructivist platform for students from the middle years to produce graphs without the need to know the exact format before starting and it also provides a random sampler, a ruler device, a tool to define a measure, and a history button to keep track of measures. These features make *TinkerPlots* an ideal software for carrying out the resampling procedures. In this software, students are able to construct their own investigation, rather than only be presented with the completed computer code by the teacher. Even without a real-time demonstration, the steps in the following section illustrate the resampling (without replacement) procedure for comparing groups without having to write instructions in a software language.

Resampling with TinkerPlots 1

The setting used for the presentation of resampling is taken from Shaughnessy et al. (2009) derived from the chapter by Gary Kadar first appearing in National Council of Teachers of Mathematics (NCTM, 2009). It is based on a classroom experiment that could be carried out in any class from Year 6 to university, which allows students to investigate the suggestion that it is easier for people to memorise meaningful words than nonsense words. Two lists of three-letter words are prepared, one a list of meaningful words and the other a list of nonsense words. These lists are randomly distributed face-down to members of the class, one to a student. Students are told to turn the sheet over and spend 30 seconds memorising the words. They then turn the sheet over and write as many words as they can remember on the back of the sheet. The statistical question is then related to a comparison of the performances of students who worked on each list of words: Is it easier to remember meaningful words than nonsense words? Stacked dot plots for each type of word can be created and means or medians calculated.

Shaughnessy et al. (2009) provide an exemplary classroom discussion that illustrates the issues that are likely to arise in introductory sessions with the topic, for example, going beyond developing "habits of mind" such as looking for patterns and variation, choosing a model, evaluating observations and reflecting on the reasonableness of conclusions (p. 84), to distinguishing between the real study and the "fake" or "hypothetical" studies that are randomised in order to draw informal inference conclusions (p. 92). They present data illustrating one class' results that can be easily entered into data cards in *TinkerPlots*. Figure 1a shows the original data set as it appears in a stack of data cards (one card on top) and in a table. Figure 1b shows a plot of the data separated by type of word and labelled with the median of the number of words remembered for each word type.

The question is, how unusual is this difference in medians (4 words) for the number of words remembered for the two word types, when there is no genuine advantage to remembering the meaningful words? The Sampler in *TinkerPlots* provides the opportunity to investigate how unusual the difference is based on random assignment without replacement of the attribute

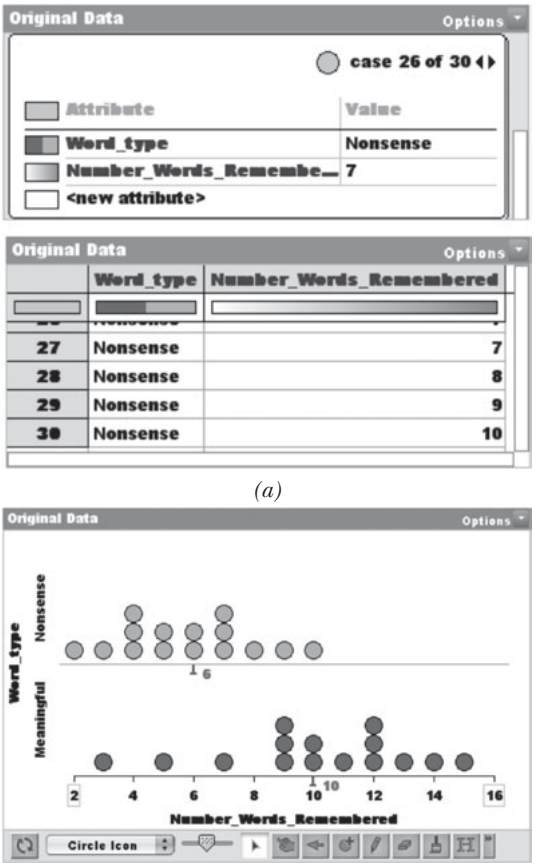


Figure 1. Basic features of TinkerPlots.

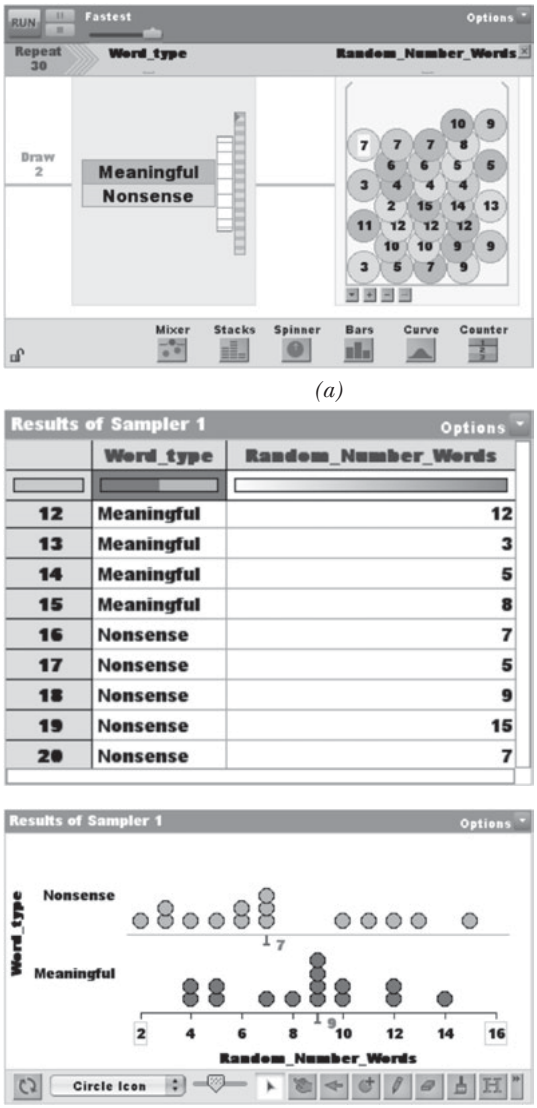
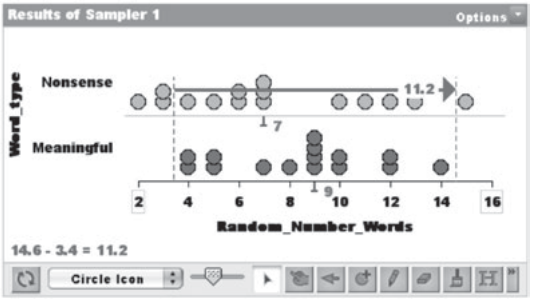


Figure 2. Creating a random resampling of the data.

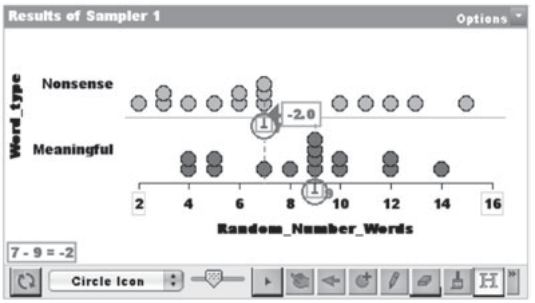
Number_Words_Remembered, hence assuming there is no genuine difference in the treatments and any differences found arose solely from the random assignment process used in the study. After being emptied of its initial content, the Sampler is set up with a Counter containing the Word_type data in the order they appeared in the data cards and a Mixer containing the Number_Words_Remembered (Figure 2a). Resampling is done from the Mixer without replacement with a run size of 30 (the original combined sample size of the two groups) and the values assigned to the two types of words (in the Counter).[2] The results appear in a new table Results of Sampler 1, now labelled Random_Number_Words, and these can be plotted as before (Figure 2b).

With the Plot window selected, the Ruler is chosen from the top tool bar (Figure 3a). The ends of the Ruler can be dragged to the medians of the two groups of word types, moving the mouse to the median symbol, which causes a circle to appear. The Ruler then measures the distance between the two medians and records this in the lower left corner of the plot. Clicking the History (H) button in the tool bar below the plot creates boxes around the medians and the equation of their difference (Figure 3b).

2 It is also possible (and equivalent) to use two Mixers and randomly sample both attributes, pairing the data. Some students find this method more intuitive.

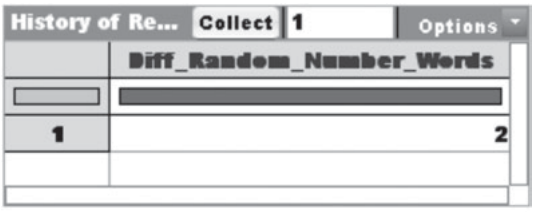


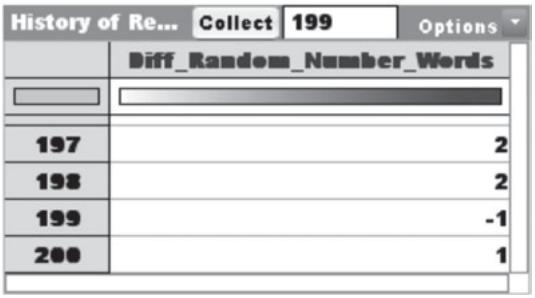
(a)



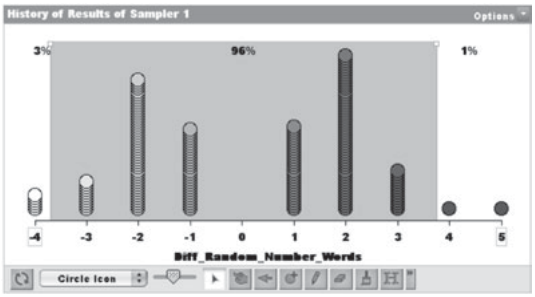
(b)

Figure 3. Using the Ruler, Measure tool and History button.





(a)



(b)

Figure 4. Collecting 200 newly resampled samples and plotting the difference of medians.

Clicking on the equation box creates a new table of the History of Results of Sampler 1. Initially it has one entry from one run of the resampling. As many resamples as desired can be added by inserting the number desired in the Collect box in the History of Results of Sampler 1 table (Figure 4a). At the same time a new Plot can be created to display the resulting differences in medians (Figure 4b), in this case 200 resampled differences.

It is then possible to observe how many of the resampled cases create a difference of four or more to compare to the original difference (or a difference of -4 or less). This is a two-tailed test also acknowledging the possibility of nonsense words being easier to remember than meaningful words. Using the Divider tool highlights that in this case there are only 4% of the 200 resampled differences as extreme or more extreme than 4, or 1% if the alternative hypothesis is that meaningful words are easier to remember. This provides evidence that there is some other reason than "chance" for the observed tendency to remember more of the meaningful words. (Students should also be able to explain the symmetry and centre of this distribution based on the underlying model of no genuine treatment effect.)

To focus attention on the importance of sample size and the ease with which technology allows this approach to be employed, Shaughnessy et al. (2009) included another data set with 60 values representing 30 students randomly assigned meaningful words and 30 randomly assigned nonsense words. The data are shown in the Plot in Figure 5, where the difference in medians for the two

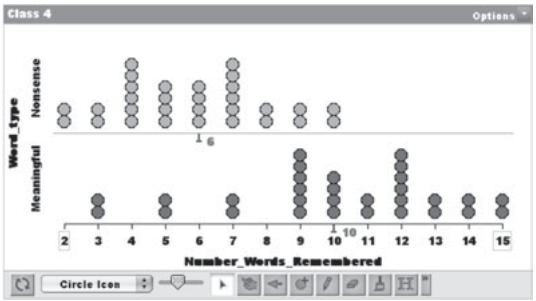
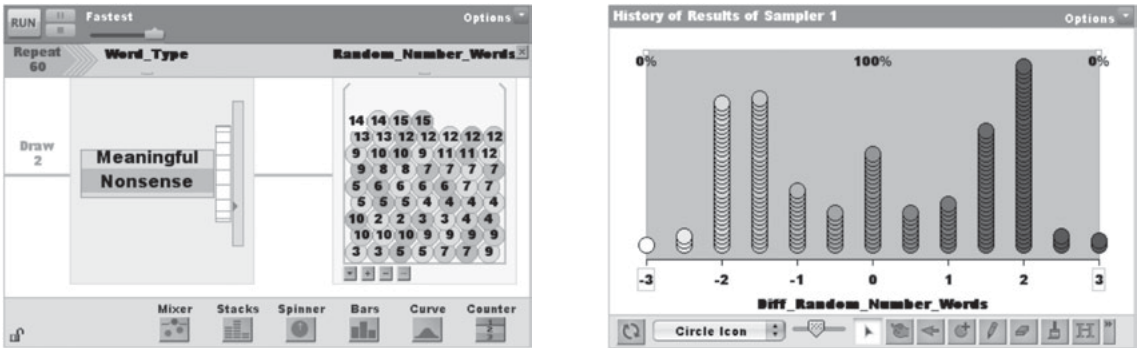


Figure 5. Data from a class of 60 students.

groups is four, the same as for the original data set of sample size 30. Shaughnessy et al. note that novice students are initially likely to expect either more variation in the distribution of differences in medians or expect no change in the significance of the results because the difference in medians is the same (p. 99).

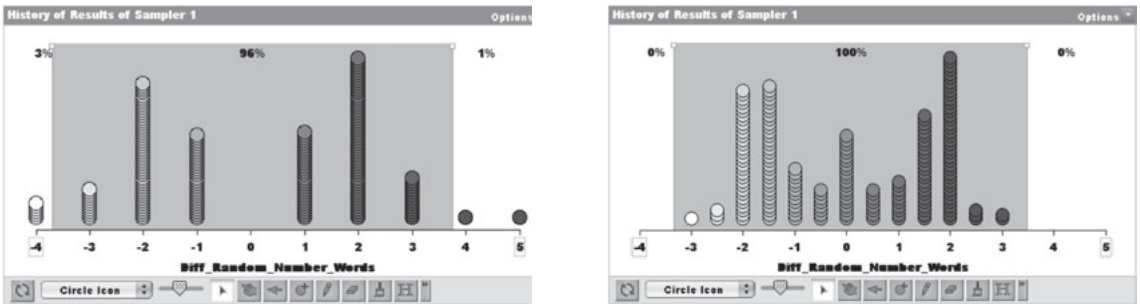
To test their conjectures, the same procedure is followed in *TinkerPlots* in placing the two attributes in an empty Sampler and drawing out 60 values from the Mixer, the number of words remembered collected randomly without replacement (Figure 6a). The process is the same as illustrated for the original data set of 30, with Figure 6b showing the plot for 200 resamples and no samples with a difference in medians as great as 4.



(a) *(b)*

Figure 6. Summary of resampling for a class of 60.

Placing the History of Results for the two Samplers on the same scales, as in Figure 7, shows the reduced variation in the differences in medians for the larger sample size. Because of the larger the sample size, there is less variability in the sample medians, and the evidence is stronger than before that the difference in memory for meaningful and nonsense words is not a chance difference. The focus for student discussion is more on the reduction in spread seen in Figure 7 rather than just that "*n* is bigger than 30."



(a) *(b)*

Figure 7. Comparing the resampling distributions for samples of size 30 (top) and 60 (bottom).

It appears safe to conclude that there is a genuine tendency for people (as represented by these students) to remember more meaningful words than nonsense words. Furthermore, because the actual study involved random assignment of students to the two conditions it is possible to claim that this as a cause-and-effect relationship.

Resampling with TinkerPlots 2

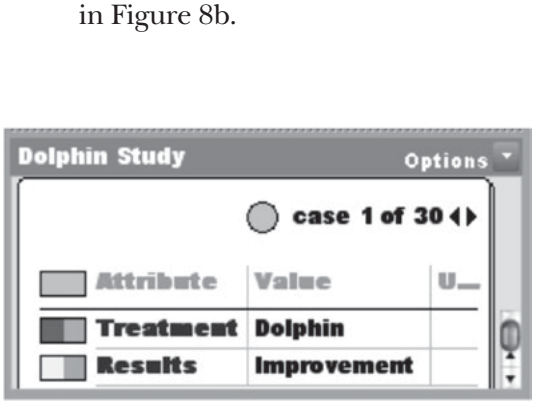
The first example presents an alternative to the traditional two-sample *t*-test for means and is a commonly used application of resampling. In fact, the ability to choose different statistics (e.g., comparing medians instead of means) provides students with a very flexible and extendable tool that can be applied much more extensively than *t* tests.

The next example is based on a two-way table that presents frequency data based on an experiment described in Rossman (2008) where 30 patients diagnosed with mild to moderate depression were randomly allocated to two treatment groups where they engaged in the same amount of time swimming and snorkelling each day for four weeks. One group, however, participated in the presence of bottlenose dolphins and the other did not. The patients had no other treatment. The results are shown in Table 1.

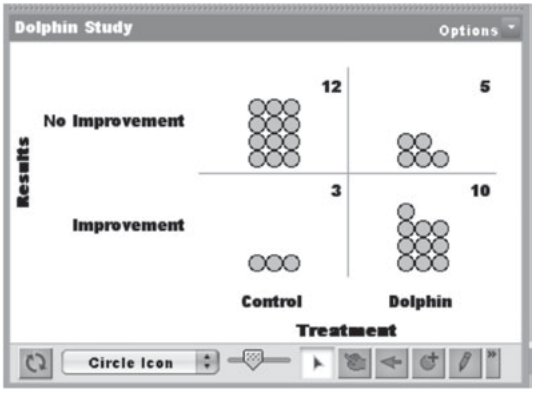
Table 1. Results of Dolphin Therapy Experiment (from Rossman, 2008).

	Control group	Dolphin Therapy	Total
Did not show substantial improvement	12	5	17
Showed substantial improvement	3	10	13
Total	15	15	30

The question arises as to whether this result, favouring the treatment including dolphins, was likely to have happened by chance alone. The process is the same as above—resampling the results using the Mixer without replacement and assigning values to the treatments in the Counter—recording the statistic from each resampling, say the number who were assessed as showing substantial improvement after swimming with the dolphins. This process mimics the original random allocation of patients but assumes 13 patients will show substantial improvement regardless of the treatment they received (hence assuming no genuine treatment effect). How do these numbers from resampling compare to the 10 improvers in the dolphin group observed in the study? The data from Table 1 can be entered into *TinkerPlots* via the Data Cards (Figure 8a) or Table with two categorical attributes: *Treatment* and *Results*. The values of the attributes are Dolphin or Control for *Treatment* and Improvement or No Improvement for *Results*. The data are seen in a two-way plot with the same information as Table 1 in Figure 8b.

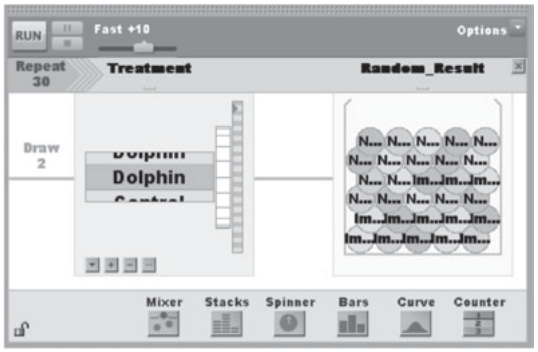


(a)

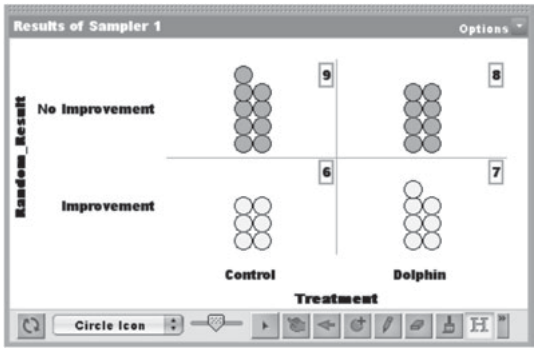


(b)

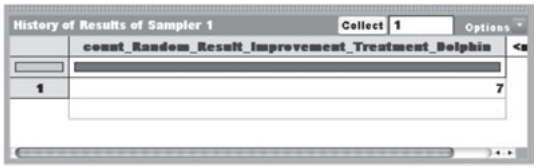
Figure 8. Data for the Swimming with Dolphins Study.



(a)

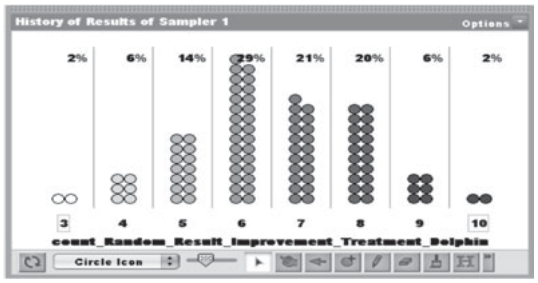


(b)

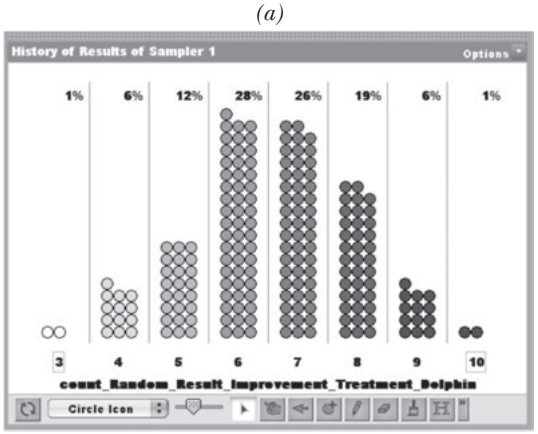


(c)

Figure 9. Setting up the resampling process for the Dolphin Study.



(a)



(b)

Figure 10. Results of resampling in the Dolphin Study.

The Sampler is set up again with the Counter and with the Mixer sampling the values of the *Results* attribute randomly without replacement (Figure 9a). For the first run the plot of the resample is shown with the count in each cell and the History button clicked to highlight the counts. Further clicking on the "7" in the Dolphin/Improvement cell collects the history of the resamples with respect to this cell (Figure 9b) to compare with the "10" that was observed in the original experiment (Rossman (2008) equivalently used the difference in conditional proportions rather than counts, which is also possible with *TinkerPlots*). The result is recorded in the History of Results of Sampler 1 table (Figure 9c).

Repeating the resampling process 100 times (Figure 10a) shows that by chance, it would be expected that approximately two times out of 100, 10 or more of the 13 subjects with substantial improvement would be observed for those swimming with dolphins. Repeating the resampling another 100 times (Figure 10b), no additional results as large as 10 (or more) occur and the estimate of the probability of the result occurring by chance reduces to 0.01. (As Rossman, 2008, reports, the exact probability is calculated as 0.0127.) Because this probability is small and the study was a randomised experiment, the evidence suggests that there is some confidence that swimming with bottleneck dolphins is more effective in improving the patients' depression than the control (swimming without bottleneck dolphins).

Although carried out in a virtual environment, the collection and display of the statistics from the resampled data can provide a "concrete" experience for students. This is especially true if they have conducted a physical random resample themselves a few times before using the software. In both of these examples, watching the process in *TinkerPlots* as plots are replaced and statistics recalculated for each resample reinforces the randomisation at work and leads to answering the question, "could the initial observed

outcome have happened by chance?" If we answer "no" or at least "very unlikely," then, because both studies utilised random allocation to treatments, we conclude that there is evidence for a causal relationship: meaningful words are easier to remember than nonsense words and swimming with bottlenose dolphins is more effective in improving patients' depression than the control.

For students who are reluctant to claim a difference in two treatment groups if the distributions are partially overlapping, watching the resampling process could help build intuitions about what is a meaningful difference. For those initially willing to declare any difference as significant, examples where the initial difference appears in the middle of the distribution of resampled differences (e.g., Stephenson et al., 2010) could help build intuitions about when there is very little evidence that a genuine difference exists. These examples can easily be extended to cases where students examine the effect of sample size and within group variability (for quantitative data). Acknowledging the need for an understanding of chance processes, the benefit of this approach for those who are beginners in the realms of inference is that the experience takes place without the baggage of theoretical assumptions and formal calculations of *p*-values while building inferential reasoning.

Place in the curriculum for resampling

Those who have used resampling as an introduction to ideas of formal inference report various degrees of positive outcome. Simon, Atkinson and Shevokas (1976) are the earliest to experiment with resampling as the Monte Carlo method. In three different classroom settings at the tertiary level they report both achievement and attitude differences favouring the randomisation procedure over the conventional theoretical procedure. Ricketts and Berry (1994) report positive feedback from their undergraduates using the *Resampling Stats* software.

> The sampling method is, in my opinion, far easier to understand than the mathematical solution … A very clear difference I thought, was that the resampling method makes one feel that we are physically doing it, or actually seeing it physically being done, without having to take any theoretical mathematics into consideration. (p. 43)

Simon et al. (1976) make it clear how they see the place of resampling methods in relation to conventional methods of teaching statistical inference. Using the terminology of resampling their comments appear relevant to the Australian context.

> Lest this be unclear or seem to equivocate: Where there is limited time, or where students will not be able to grasp conventional methods firmly, we advocate teaching the [resampling] approach, and perhaps that only. Where there is more time, and where students will be able to well learn conventional methods, we advocate (a) teaching [resampling] methods at the very beginning as

> an introduction to statistical thinking and practice; and (b) afterwards teaching the [resampling] method with the conventional method as alternatives to the same problem, to help students learn analytic methods and to give them an alternative tool for their use. (p. 15)

As Simon et al. (1976) go on to note, this method involves students interacting with the pedagogy and the data, developing intuitions about what distributions look like and what it means to be "statistically significant." Once the randomisation process is accepted, there is no need to take anything else on faith, as is usually the case for the analytical methods associated with traditional hypothesis testing. Tintle et al. (2011) more recently report significantly better outcomes on test items assessing statistical inference for tertiary students taught a randomisation-based curriculum (modelled on Rossman and Chance (2008)) compared to a more traditional curriculum. The examples presented by Schaeffer and Tabor (2008) and Shaughnessy et al. (2009) are specifically designed for secondary students and the draft US Common Core Mathematics Curriculum (Common Core State Standards Initiative, 2010, p. 82) includes randomised experiments and resampling ideas (S-IC.5) for the high school.

Given the potentially crowded nature of the *Australian Curriculum: Mathematics* at Years 11 and 12, one could suggest the advice of Simon et al. (1976) is relevant to building intuitive ideas about inference. Cobb (2007) presents an even stronger argument than Simon et al. for changing the face of introductory statistics courses. Building on the changes brought about by computers to automate calculations and graphics by the end of the 20th century, Cobb goes one step further.

> Just as computers have freed us to analyse 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 randomised 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. (pp. 1–2)

Using as his analogy of how Copernicus changed the view of the earth as the centre of the universe, Cobb would throw away the normal approximation to a sampling distribution as the centre of the statistics and replace it with the core logic of inference: the three Rs. Cobb's three Rs are (1) *randomise* data production, (2) *repeat* by simulation to see what is typical, and (3) *reject* any model that puts the data in its tail. In relation to building students' intuitions about inference, providing a few examples of resampling in this manner would appear to provide the evidence required by Makar and Rubin (2009) to measure the degree of uncertainty with which one is able to generalise from the original sample of data collected, the evidence being strongest when the original data meet random criteria of selection or assignment. Why not bite the bullet and give students in Years 11 and 12 the chance to extend their previous work with exploratory data analysis in Years F–10 in a meaningful way to build an understanding of the logic of inference? If they go on to traditional

statistics courses at university, they will be able to apply the logic in the traditional environment.

Given the step-by-step visual presentation of the process of resampling that is created through the use of *TinkerPlots*, it is even possible to make a more radical suggestion. Why not make resampling available to students before Year 11? Relative to the traditional theoretical approach, the resampling process could begin with tactile simulations using concrete materials and it could be introduced as soon as students are experiencing simulations in the curriculum. Computer simulations are suggested in relation to Chance as early as Year 6 (ACARA, 2011, p. 32) and the TIMES Project (2011) uses simulation to illustrate sampling variation for a known proportion of a population with two outcomes in Year 8. This reflects the focus on samples and populations that develops across the middle years. Given that research has shown that students in Years 5 to 7 can create and interpret scatterplots (e.g., Fitzallen & Watson, 2011; Watson & Donne, 2009), which are not explicitly mentioned in the mathematics curriculum until Year 10 (ACARA, 2011, p. 46), potentially many students in Year 10 and earlier could develop intuitions about informal inference given straightforward examples as in Shaughnessy et al. (2009) or Rossman (2008) and software as creative as *TinkerPlots*.

References

Arnholt, A. T. (2007). Resampling with R. *Teaching Statistics, 29*(1), 21-26.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2011). *The Australian Curriculum: Mathematics, Version 1.2, 8 March 2011.* Sydney, NSW: ACARA. Retrieved from http://www.australiancurriculum.edu.au/

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics* (CD-ROM), Salvador, Bahia, Brazil, 2–7 July, 2006. Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D., Gil, E., & Apel, N. (2007, August). *What is hidden behind the data? Helping young students to reason and argue about some wider universe.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

Christie, D. (2004). Resampling with Excel. *Teaching Statistics, 26*(1), 9–14.

Clements, C., Erickson, T., & Finzer, B. (2007). *Exploring statistics with Fathom Dynamic Data Software Version 2.* Emeryville, CA: Key Curriculum Press and Key College Press.

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

Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics.* Retrieved from http://www.corestandards.org/the-standards

Engel, J. (2010). On teaching bootstrap confidence intervals. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute. Retrieved from www.stat.auckland.ac.nz/~iase/publications.php

Fitzallen, N., & Watson, J. (2011). Graph creation and interpretation: Putting skills and context together. In Clark, J., Kissane, B., Mousley, J., Spencer, T., & Thornton, S. (Eds.), *Mathematics: Traditions and [new] practices* (Proceedings of the AAMT/MERGA conference, pp. 253–260). Adelaide, SA: AAMT and MERGA.

Johnson, R. W. (2001). An introduction to the bootstrap. *Teaching Statistics, 23*(2), 49–54.

Key Curriculum Press. (2005). *Fathom. Dynamic Data Software, Version 2.01.* [Computer software]. Emeryville, CA: KCP Technologies.

Konold, C. & Miller, C. D. (2005). *TinkerPlots: Dynamic data exploration* [Computer software]. Emeryville, CA: Key Curriculum Press.

Konold, C. & Miller, C. D. (2011). *TinkerPlots: Dynamic data exploration* [Computer software, Version 2.0]. Emeryville, CA: Key Curriculum Press.

Makar, K., & Ben-Zvi, D. (2011). The role of context in developing reasoning about informal statistical inference. *Mathematical Thinking and Learning, 13*(1–2), 1–4.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82–105.

National Council of Teachers of Mathematics. (2009). *Focus in high school mathematics: Reasoning and sense making*. Reston, VA: Author.

Park, J., delMas, R. C., Zieffler, A., & Garfield, J. (2011). A research-based statistics course for tertiary students. In *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*. Dublin: ISI. Retrieved from http://isi2011.congressplanner.eu/pdfs/450364.pdf

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning, 13*(1–2), 27–46.

R Development Core Team. (2004). *R: A language and environment for statistical computing*. [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Reaburn, R. (2011). *Students' understanding of statistical inference: Implications for teaching*. Unpublished PhD dissertation, University of Tasmania, Hobart.

Ricketts, C., & Berry, J. (1994). Teaching statistics through resampling. *Teaching Statistics, 16*(2), 41–44.

Rossman, A. J. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal, 7*(2), 5–19. Retrieved from http://www.stat.auckland.ac.nz/serj

Rossman, A., & Chance, B. (2008). *Concepts of statistical inference: A randomization-based curriculum*. San Luis Obispo, CA; California Polytecnic University. Retrieved from http://statweb.calpoly.edu/csi

Scheaffer, R., & Tabor, J. (2008). Statistics in the high school mathematics curriculum: Building sound reasoning under uncertain conditions. *Mathematics Teacher, 102*(1), 56–61.

Shaughnessy, J. M., Chance, B., & Kranendonk, H. (2009). *Focus in high school mathematics reasoning and sense making: Statistics and probability*. Reston, VA: National Council of Teachers of Mathematics.

Simon, J. L. (1997). *Resampling: The new statistics*. (2nd ed.). Arlington, VA: Resampling Stats Inc.

Simon, J. L., Atkinson, D. T., & Shevokas, C. (1976). Probability and statistics: experimental results of a radically different teaching method. *American Mathematical Monthly, 83*, 733–739.

Stephenson, W. R., Froelich, A. G., & Duckworth, W. M. (2010). Using resampling to compare two proportions. *Teaching Statistics, 32*(3), 66–71.

Taffe, J., & Garnham, N. (1996). Resampling, the bookstrap and Minitab. *Teaching Statistics, 18*(1), 24–5.

TIMES Project. (2011, June). *Data investigation and interpretation: A guide for teachers – Year 8*. Melbourne: University of Melbourne for the International Centre of Excellence for Education in Mathematics. Retrieved from http://www.amsi.org.au/teacher_modules/Data_Investigation_and_interpretation8.html

Tintle, N., VanderStoep, J., Holmes, V-L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1). Retrieved from http://www.amstat.org/publications/jse/v19n1/tintle.pdf

Watson, J. M. (2008). Exploring beginning inference with novice grade 7 students. *Statistics Education Research Journal, 7*(2), 59–82.

Watson, J. M., & Donne, J. (2009). *TinkerPlots* as a research tool to explore student understanding. *Technology Innovations in Statistics Education, 3*(1), 1–35. Retrieved from http://repositories.cdlib.org/uclastat/cts/tise/vol3/iss1/art1/

Watson, J., & Wright, S. (2008). Building informal inference with *TinkerPlots* in a measurement context. *Australian Mathematics Teacher, 64*(4), 31–40.

Wood, M. (2004). Statistical inference using bootstrap confidence intervals. *Significance, 1*(4), 180–182.