ACCEPTING THE CHALLENGE: CONSTRUCTING A RANDOMISATION PATHWAY FOR INFERENCE INTO OUR TRADITIONAL INTRODUCTORY COURSE

A. Marie Fitch and M. Regan
The University of Auckland, New Zealand
m.fitch@auckland.ac.nz

Current thinking in the statistical education community is that a randomisation pathway, with the aid of dynamic visualisations, will provide students with a more accessible and better conceptual understanding of the thinking underpinning statistical inference than a traditional normality-based approach. We report on where we are currently at in introducing randomisation methods into our large introductory course. Continuous reflection on our construction of this pathway has been and will continue to be an important part of working out its implementation. We include an outline of some principles we have tried to adhere to, issues we have encountered, and the main constraints by which we have been bound.

INTRODUCTION

In 2007 Cobb laid down a challenge to the statistics education community to re-consider how statistical inference was being introduced to students in the traditional introductory course. He stated that "we need a new curriculum, centered not on the normal distribution, but on the logic of inference" (Cobb, 2007). We have seen the challenge accepted and acted upon by textbook authors and course leaders (e.g., Lock et al., 2013; Rossman, 2008; Tintle et al., in press; Watkins et al., 2011; Catalysts for Change, 2012) through underpinning their introductory courses with randomisation methods.

We have also acted upon the challenge by using randomisation methods for introducing inference into our large introductory statistics course. The end of 2012 saw the culmination of a five-year research project led by Maxine Pfannkuch, Chris Wild and Pip Arnold (University of Auckland) on the development of a conceptual pathway for learning inference across the New Zealand (NZ) secondary school statistics curriculum and into the standard undergraduate introductory statistics course (see Pfannkuch and Wild, 2012). In 2011 we began developing teaching and learning resource materials for re-sampling methods and Chris Wild and his team began work on creating dynamic visualisation software for learning and analysis (see *VIT* – visual inference tools – http://www.stat.auckland.ac.nz/~wild/iNZight/downloads.html). Then in 2012 we began trialling these resources in some Year 13 classes throughout NZ and in introductory statistics courses in two NZ universities. We continued trialling our materials in our first year introductory statistics courses throughout 2013. Using what we learnt over those first two years together with initial findings from some of the most recent research led by Maxine Pfannkuch and Stephanie Budgett (Budgett et al., 2013) and from the reported experiences of others (Tintle et al., 2011 and 2012; Holcomb et al., 2010), we began some major revisions with the aim of having them incorporated for use in 2014.

In this paper we report on some of our experiences thus far in constructing a randomisation pathway for inference into our large introductory course first year university course.

CONTEXT AND CONSTRAINTS

Our introductory statistics course is the largest course at the University of Auckland with an enrolment of over 4500 students per year. For a number of years the course has been delivered by a core team of 5 lecturers with assistance from a designated course administrator and a large team of student tutors. All lecturers use a shared set of teaching and presentation resources which have been developed over a number of years. Weekly team meetings throughout each semester allow for systematic evaluation of course content and teaching resources. We enjoy the advantage of the collected experience and expertise within the team but a big disadvantage is that such a large and complex operation requires a long lead time for making even the smallest of changes. A lecture comprises between 250 and 550 students. Classes of this size also bring with them their own delivery constraints. Our introductory course serves as service course for a number of client

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]

departments. We are obliged to meet the demands of these client departments and therefore cannot use re-sampling methods only but must also include the traditional normality-based methods.

COURSE DESIGN

We take two bites, which we will call "the first bite" and "the second bite", at introducing hypothesis testing into our course. In the second week of the course we take the first bite which uses the randomisation method only with completely randomised experiments involving two or more groups for both quantitative and qualitative data. In the second bite, mid-way through our 12 week course, we re-visit the randomisation method for inference using a formal hypothesis testing structure with conventional language and terminology and extend it to cover observational studies. We randomise the *t*-statistic as a way of introducing the Student's *t*-distribution and the traditional normality-based *t*-test.

The First Bite: Some Principles

We try to grab the students' attention in the course by getting them into some real and interesting statistics as early as possible. In the second week of lectures we introduce the randomisation test. Our overall goal with this first exposure to the randomisation method is to help students grasp and understand what Cobb (2007) describes as "the core logic of inference". We do not use a formal hypothesis testing structure or any of its associated language but instead we appeal to an almost intuitive form of argumentation. We try to use language which is linked to or reflects the concept, e.g., using "the re-randomisation distribution" instead of the "randomisation distribution" and using "tail proportion" instead of "*P-value*". We postpone dealing with important but nevertheless distracting issues such as parameters and what they represent and the 1-tail and 2-tail issue. We focus on how the data were produced and the logic of the argumentation. Key concepts highlighted are *chance acting alone*, the explanation that *the observed difference just happened by chance acting alone* and how to assess the implausibility of this explanation by using the data to *simulate what happens under chance alone*. At this stage we use only completely randomised experiments, thus avoiding observational studies, so that the simulation process closely mimics the data production. We confine ourselves to what happened in the actual study and we are careful in language usage to avoid implying any generalisations beyond the participants in the study, i.e., our focus is restricted to the (statistical) significance of the observed difference. In summary, we make every effort to *reduce cognitive load* by minimising the number of concepts presented in this first encounter and concentrating our attention on the goal of grasping the logic underlying inference.

The First Bite: Approach

We start with an adaptation of the widely known finger tapping rates–caffeine experiment (Hand, 1996). We use only two groups (caffeine and no-caffeine) with five made-up observations per group and we compare medians. The observed difference the two group medians is 4 taps per minute. (Within two lectures we return to this experiment using the actual data which comprises three groups with 10 observations per group.) Only five observations and medians, rather than means, are used initially because the median value can be *visually* and quickly obtained. With up to 500 students participating in a hands-on activity, we need to minimise busy distractions such as calculating the mean value. We also believe the argument is more compelling if the first exercise demonstrates a statistically significant result.

Prior to starting this example we have already demonstrated visually, using *VIT,* that random allocation alone can generate an apparent difference between two groups. Students are then more likely to understand what is meant by the explanation that the observed difference, 4 taps per minute, could have just been due to chance acting *alone*. The argument we present is along the following lines: We ask "Is chance alone likely to generate differences as big as 4 taps per minute?". We answer this question by *seeing* what's *likely* and what's *unlikely* under chance alone by simulating the data production through re-randomisation under chance alone, beginning with a ticket-tearing hands-on activity and then switching to *VIT*. Under chance alone, we see that the differences between the medians are almost always smaller than 4 taps per minute and we conclude that we are *pretty sure* that chance was not acting alone in the actual study. This then raises the

question if chance was not acting alone, then what else was acting to produce the observed difference between the medians of 4 taps per minute? We remind our students that the participants were randomly allocated to one of two groups and each group received a different treatment and that this allows us to classify all explanations other than the difference-in-treatment explanation as "chance explanations". This means that we are pretty sure that the caffeine, together with an element of chance, produced the observed difference between the medians of 4 taps per minute. We thus make *a claim* that the caffeine had an effect on the median finger tapping rate in this study.

We follow up with three more two-group exercises from a study in which researchers Attwood et al. (2012) were interested in determining whether glass shape (curved or straight) influences consumption rates (in minutes) for alcoholic beverages. In the first of these exercises (large glass, alcoholic beverage) we find that the observed difference between two group means is so large that it is unlikely to have happened by chance alone and we conclude that, in the actual study, the shape of the glass had an effect on the mean time taken to drink the alcoholic beverage. In the remaining two exercises, large glass/non-alcoholic beverage and small glass/alcoholic beverage, we find that under chance alone it is not unusual to get differences between the means as big or bigger than the observed differences and thus we have no evidence against the chance alone explanation. We conclude that the observed differences were the result of either just chance acting alone or the glass shape together with an element of chance but that we do not have enough information to make a call as to which.

We return to the finger-tapping rate/caffeine study and apply the same argument to a three group situation. The statistic of interest is the average absolute deviation from the overall mean. With *VIT* we can see that because the average deviation is so large it is unlikely to have happened by chance alone.

Two exercises involving qualitative data follow, one a two-group comparison and the other a three-group comparison. We highlight that there is nothing new in the way we argue in these new contexts, the logic of argument is exactly the same as in our first exercise.

The First Bite: Some Teaching Recommendations

1. Use hands-on simulations before introducing computer simulations. The randomisation method lends itself to tactile and visual experiences.
2. Use software which presents visualisations of simulations and take time to explicitly link the software visualisation and the hands-on activity so that students understand what the software visualisations are showing. See Recommendation 4. below.
3. Demonstrate what is meant by "chance alone". e.g., use the "Randomisation Variation" module in *VIT* to show visually how random allocation alone can generate apparent differences. Budgett et al. (2013) reported that some students have difficulty in understanding the concept of "chance alone".
4. Distinguish clearly between the observed data and the data simulated under chance alone. Preliminary findings from Gould et al. (2010) reported that a common misconception held by students was that the re-randomisation distribution was the "real" distribution.

The Second Bite

When we return to testing mid-way through the course we try to leverage off the gains made in the first bite in which we focused on the logic of the argument underpinning inference through testing. We now introduce fundamental ideas about hypothesis testing in a more formal structure, extend to situations which allow sample-to-population inferences and introduce the *t*-test. We confine ourselves to situations which involve only the difference between two means or the difference between two proportions and with null hypotheses of the "it makes no difference" variety.

The Second Bite: Some Principles

One of our goals is to get students to always think about the data production process and clearly identify any randomisation involved. In the first bite we purposefully dealt only with situations in which the data were produced through random allocation of units to treatment groups

(experiments) and thus significant conclusions in the exercises were all of the same "experiment-to-causation" type inference. We now introduce situations in which the data have been produced through random sampling of units from two populations for which we make "sample-to-population" type inferences. There is a need to repeatedly and clearly remind students that the defining of the parameters (see *The second bite: issues*) and the type of inference made in the conclusions are determined by how the data are produced, and in particular, the form of randomisation (if any) in that data production process.

At this stage we get the students to behave more as "learners" than practising statisticians. We make them conduct the hypothesis tests "by hand". This means that we want them to show explicitly all the steps involved in order to gain a better understanding of the testing method. The algorithm for conducting a test is presented as:

hypotheses → data → test statistic → *P-value* → interpretation.

Through providing the students with a step-by-step template for all hypothesis tests conducted by hand we reinforce that this algorithm is the same irrespective of the particular test used.

Language and terminology becomes more formal and conventional than in the first bite. For example, we begin by using "hypothesis to test" instead of "statement to test" and we replace "tail proportion" with "*P-value*".

The Second Bite: Issues

As soon as we move to a formal hypothesis test structure we face the issue of parameter use and definition. Parameter definition is straight forward in a sample-to-population context. For experiments, it has been common practice to define parameters in an introductory course in terms of "imaginary" populations. This is no longer acceptable. The defining of parameters is directly linked to the data production which in turn determines the scope of the inference. For an experiment with a completely randomised design, we define the parameters in terms of "if all participants had been in group X" which is not as obvious as when working in a sample-to-population context. For this reason we decided to introduce formal hypotheses in a sample-to-population inference setting first.

This means we need to be able to justify the use of the randomisation test in a sampling situation and then explain what "*chance acting alone"* means in a sampling situation. We conjecture that it will seem reasonable to students that "if the populations, and hence the means, are the same, then observations could have come from either population and hence appear in either sample". We tighten our language later in the course using "equality of distributions" rather than "equality of populations" when we consider assumptions underlying the randomisation test. In a sampling situation, "chance acting alone" equates to "sampling variation". When the null hypothesis (of equal means) is true, the observed difference is a manifestation of sampling variation, that is, it is only due to chance.

The next issue we confronted was how to introduce the *t*-test. Underlying theory aside, there are two key differences between a *t*-test and a randomisation test: which test statistic is used and how its null sampling distribution is approximated. In a *t*-test, the test statistic for no difference between two means is a scaled difference between the means (the *t*-statistic) and a Student's *t*-distribution is used to approximate its null sampling distribution. In a randomisation test we may re-randomise any statistic and the resulting re-randomisation distribution is used to approximate its null sampling distribution. We realised that first we have to introduce the concept of a test statistic, a measure of the discrepancy between what we see in the data and what we would expect to see if the null hypothesis were true. We then show that, for the same data, re-randomising the *t*-statistic yields a very similar *P-value* to re-randomising the "natural" test statistic of choice, the actual difference between the means. We then observe that this re-randomisation distribution of the *t*-statistic is approximately a Student's *t*-distribution. An (approximate) *P-value* can be obtained using this theoretical distribution (the *t*-test).

The final issue is how to introduce and justify the use of *t*-procedures in an experimental situation where a randomisation test is the "gold standard". We decided not to try to justify it other than to tell students that using the *t*-test in an experimental context is very common in practice.

The Second Bite: Approach

The second bite begins with revision of our earlier introduction to testing. We conduct a randomisation test in the context of a randomised experiment again focusing the students' attention on the underlying logic of the argument. We then embark on presenting a more formal structure using conventional language and terminology and shift from experiment-to-causation inference to sample-to-population inference through the following example.

We use a subset of some pizza size data (Dunn, 2012) to justify a claim that Eagle Boys Hawaiian thin-crust pizzas are larger than Domino's. The concept of two competing hypotheses is introduced and the hypotheses are written formally in terms of the difference between two means, with the alternative being 1-sided. At this stage there is a discussion about "1-sided" and "2-sided" alternative hypotheses. The "natural" test statistic to use in this case is the raw difference between the estimate and the hypothesised value (0) which has the same value as the estimate $\overline{x}_{\text{EB}} - \overline{x}_{\text{D}}$. *VIT* is used to construct a re-randomisation distribution of the test statistic by randomly re-assigning the data to the two samples 1000 times (a randomisation test). A somewhat informal justification for doing this random re-assignment is given. We look at the tail proportion, as we did in the first bite, and we use this tail proportion to estimate the *P-value*. This (estimated) *P-value* is interpreted in terms of strength of evidence against the null hypothesis in favour of the alternative hypothesis. We also show how to use the *P-value* in a decision making process (significant or not, i.e., reject the null hypothesis or not). Finally we raise the "statistical significance versus practical significance" issue by asking if the actual difference in size would be worthy of consideration when choosing between the two brands.

Having set up the step-by-step template, we then repeat this example using a *t*-statistic (a scaled difference) as the test statistic. Re-randomising the *t*-statistic gives an estimate for the *P-value* very similar to the previous *P-value*. We show visually that the re-randomisation distribution of the *t*-statistic can be approximated, under certain conditions, by a theoretical distribution called a Student's *t*-distribution. This means that we obtain very similar estimates for the *P-value* in all three cases: re-randomising the natural test statistic of choice ($\overline{x}_{\text{EB}} - \overline{x}_{\text{D}}$), re-randomising the *t*-statistic, and using the theoretical Student's *t*-distribution. In summary, we show that the *P-value* can be estimated through simulation as we did above (the randomisation test) or theoretically by using the Student's *t*-distribution (the *t*-test). Several examples using a *t*-test for no difference in both means and proportions covering both 1-tailed and 2-tailed tests follow.

We return to experiment-to-causation inference using a randomisation test in a formal hypothesis testing framework. We finish off this second bite with an example using the *t*-test in the context of a randomised experiment.

CONCLUSION

In 2012 we started using randomisation methods in our introductory course because we believe that these methods will make it easier for students to understand the logic underlying statistical inference. By the end of the 2nd week of the course we had introduced, with the aid of specially designed visualisation software, the randomisation method for examples involving completely randomised designed experiments with two treatment groups. It was a small change but nevertheless our first step towards a future introductory course which eventually we hope will be underpinned with randomisation and bootstrapping methods only. Positive student feedback and initial findings from research led by Pfannkuch and Budgett (Budgett et al., 2013) over this two-year trial period have been encouraging.

We have made more changes which will be trialled in 2014. We have extended the first introduction to include experiments with more than two treatment groups for qualitative as well as quantitative data. Also, later in the course is our first attempt at marrying the randomisation method and the traditional normality-based approach. We use a randomisation test for sample-to-population inference and we use this context to introduce the traditional *t*-test. We have not found designing this part of the pathway easy. It has challenged our own thinking and reasoning. The findings and experiences of others who have already designed introductory courses involving both randomisation and normality-based methods have helped us. There are still some issues in this area

which remain unresolved for us, for example, how to justify the use of the *t*-test in an experimental situation.

Throughout 2014 the teaching team will reflect on these latest changes and with consensus make decisions for further changes. We will also need to think about "where to from here?" including how we are going to present randomisation methods in a wider context of formal testing such as with one sample and more than two sample/group situations and in regression.

REFERENCES

Attwood, A., Scott-Samuel, N., Stothart, G., & Munafo, M. (2012). Glass shape influences consumption rate for alcoholic beverages. *PLOS ONE*.

Budgett, S., Pfannkuch, M., Regan, M., & Wild C. J. (2013). Dynamic visualizations and the randomization test. *Technology Innovation in Statistics Education, 7*(2), 1-21.

Catalysts for Change (2012). *Statistical thinking: A simulation approach to modeling uncertainty*. Minneapolis, MN: Catalyst Press.

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

Dunn, P. (2012). Pizza size data. *Journal of Statistics Education, 20*(1).

Gould, R., Davis, 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.

Hand, D., Daly, F., Lunn, A., McConway, K., & Ostrowski, E. (1996). *A handbook of small data sets*. London: Chapman and Hall

Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomisation tests. 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.

Lock, R., Lock, P.F., Lock Morgan, K., Lock, E., & Lock, D. (2013) *Statistics: Unlocking the power of data.* Hoboken, NJ: Wiley.

Pfannkuch, M., Regan, M., Wild, C., Budgett, S., Forbes, S., Harraway, J., & Parsonage, R. (2011). Inference and the introductory statistics course. *International Journal of Mathematical Education in Science and Technology, 42*(7), 903-913.

Pfannkuch, M., & Wild, C. J. (2012). Laying foundations for statistical inference. *Proceedings of the 12th International Congress on Mathematics Education, Regular Lectures 1-9, 8-15 July, Seoul, Korea* (pp. 317 – 329). [USB] Seoul, Korea: ICME-12. Online: http://icme12.org/

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

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & Vanderstoep, J. (in press). *Introduction to statistical investigations*. Wiley.

Tintle, N., Topliff, K., Vanderstoep, J., Holmes, V., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomisation-based introductory statistics curriculum. *Statistics Education Research Journal, 11*(1), 21-40.

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

Watkins, A., Scheaffer, R., & Cobb, G. (2011) *Statistics: From data to decision* (Second Edition). Hoboken, NJ: Wiley.