22 April 2013

Next Steps in Accessible Conceptions of Statistical Inference: Pulling ourselves up by the bootstraps

Chris J. Wild, Maxine Pfannkuch, Matt Regan and Ross Parsonage

The University of Auckland, Auckland, NEW ZEALAND

Abstract

With the rapid, on-going expansions in the world of data we need to devise ways of getting students much further, much faster. Working out how to do this without losing some of the things that statisticians value most dearly requires community involvement and debate. We introduce and model ideas of principled argument and conceptual analysis as a means for facilitating such debate. We then turn our attention to applying sets of principles developed early in the paper to the problem of introducing statistical inference. Bootstrapping and randomisation tests conveyed through dynamic visualisation are developed as a means of reducing cognitive demands and increasing the speed with which application areas can be opened up. We also discuss the design of software developed to enable this approach.

Keywords: cognition; graphics; resampling methods; statistical inference; statistics education; visualisation

Examples of dynamic content http://www.stat.auckland.ac.nz/~wild/TEMP/wild-paper/

1. INTRODUCTION

1.1 It's all too slow

We are all now well aware that the world of data is growing explosively – in volume, in the areas it reaches into, in how it is constituted, and in what you can do with it. By comparison, changes in what students experience are glacial. The chasm between possibility and awareness yawns ever wider. It can still take until the dying stages of a university intro-stat course to get much passed the *t*-test! Student exposure to the brave new world is likely to be, at best, quotations from hyped media stories. But how much more convincing and motivating would personal experiences of excitement with and through data be? In the meantime systems like Tableau (www.tableausoftware.com) are beginning to place powerful, accessible, visually-based data-assembling and analysis tools right into the hands of primary decision-makers themselves and Leland Wilknson's new AdviseStat (https://adviseanalytics.com/advisestat) looks remarkably close to being a real statistical expert system.

The biggest challenges that statistics education has to confront stem from our current approaches being far too slow (and also often ineffective; see Wild et al., 2010, 2011 for references). We need to devise ways to get students further, faster and with better understanding. Creating the space in which to make substantial progress involves not only jettisoning the inessential and outdated, but also more efficiently conveying the critical. But how can we do this without throwing out too many babies when we throw out the bath water?

Pre-requisite activities for substantial progress include:

- debating values and priorities to arrive at more relevant and ambitious goals
- generating new ideas about how to do things differently
- building enabling technology
- conducting student and teacher- facing research on capabilities and how people learn in this new environment
- devising enabling pedagogy
- conducting student and teacher-facing research on effectiveness of implementations of new pedagogy

Juxtaposing the urgency of need with the fact that most of us can only cope with incremental change, it has rapidly become clear to us that these things need to be addressed in parallel and with extensive community-dialogue.

The goal for early exposures to statistics that our group is beginning to prioritise most highly is the creation of excitement about "*what I can do with data and what data can do for me*". If we can only create in students a passion for data and its possibilities for their lives ("What's in it for me?") the seeds of desire for further study and lifelong learning will have been sown. For most, this probably requires creating a broad awareness and as many "aha!" and "wow!" moments as possible from multivariate data in a short space of time. (What you can do in a 12 week semester, for example, is extremely tightly constrained.) The alternative to breadth of exposure is to create an "aha" moment about something that touches someone very deeply but this differs too much from person to person to be a viable basis for a general strategy.

Bearing in mind our desire for maximally accessible approaches, we believe that technology, and more particularly technology-enabled visualisation, is the key to much of this. Technology in this conception is not about acquiring technological skills but using technology to finesse away as much as possible of all that creates distance between questions, data, the inklings of an idea and "aha!". "Visualisation", said Martin Wattenberg "is a gateway drug to statistics … People who look at visualisations will start asking statistically important questions … even without knowing the jargon" (Aldous, 2011).

Any consideration of goals must, of course, consider audience. The thinking that produced this paper was geared mainly towards the levels of exposure that usually come in the last year at high school and in the first one or two university service/applied statistics courses but should also be of value for graduate research methods courses (Wild et al., 2011, addressed lower-level exposures).

In this context we see two main structural barriers to getting much further into data analysis much faster. The first barrier is the complexity of software and the size of the knowledge base needed to use it for all but the simplest of problems. The second, and closely related to the first, is the time it takes to build up the ideas of normal-theory based statistical inference. Our group has been mounting an attack on these barriers by building both enabling technology and related pedagogy. The iNZight software project (www.stat.auckland.ac.nz/~wild/iNZight/) aims to provide an environment that allows even beginners to explore multivariate data rapidly and with a minimal learning curve. The VIT ("Visual Inference Tools", www.stat.auckland.ac.nz/~wild/VIT/) software project is aimed at providing visualisation tools to aid in the development of the core concepts of classical statistical inference using bootstrapping and randomisation tests. It is this second project that is most relevant to this paper.

Regardless of whether you agree with the case laid out above or the arguments of Section 1.2 on "road blocks", most of what follows should also be of interest to anyone who wants to teach key

ideas of inference for immediate use without mathematics, or to convey intuitive understandings of "what it is all about" prior to mathematizing.

1.2 Road Blocks

The road blocks this paper aims to help dismantle are those set up by the normal-theory based approaches to statistical inference which takes up such a large part of most introductory courses. But why should we not, in a pursuit of faster progress, just abandon inference and confine ourselves to exploratory data analysis using graphics and visualisation? The big danger in people getting their seminal experiences of data visualisation from outside statistics is that they come stripped of their statistical "safe sex" messages – that seeing should not necessarily be believing; that visualisations can be stunning, convey their messages powerfully, persuasively, memorably, and be wrong. Students need to learn that appearances can be deceiving, that what we see in data is never quite the way it really is, that it is absolutely essential that we pay attention to the stories behind the data and not just to the stories in the data. Students need to experience the on-going struggle to distinguish fact from artefact. Why is it that what we see in data is never quite the way it really is? There are four main causes: (i) systematically biased observational processes (due to distorting filters sitting between us and the reality we want to see, filtering the stream of data that we get to see); (ii) random factors (of which the most easily understood are due to sampling and randomisation); (iii) confounding (and concomitant difficulties in reaching causal conclusions); and (iv) deficiencies in the mental models within which we interpret what we see. Classical inference may only attempt to address the second of these causes but it is still very important that we deal not only with patterns but also with the uncertainties around them.

Unfortunately, the "traditional road to statistical knowledge is blocked for most," as Efron and Tibshirani (1993) have said, "by a formidable wall of mathematics." George Cobb (2007) characterises the usual introductory statistics course curriculum as Ptolemaic – an allusion to the unnecessary complexity in Ptolemy's cosmology stemming from putting the earth at the centre of his system instead of the sun. "What we teach", said Cobb, "is largely the technical machinery of numerical approximations based on the normal distribution and its many subsidiary cogs." Because of computational power this is no longer necessary or desirable in view of the conceptually simpler alternatives. "These days", he said, "we have no excuse." Even at the end of the process there are a very limited number of parameters we can make inferences about this way. In stark contrast, as soon as the basic bootstrap idea is understood to the point where it becomes compelling we can suddenly start putting confidence intervals on almost everything and even displaying uncertainties about complex features like scatterplot smooths (by supplementing a smooth with a nest of bootstrap smooths, cf. Hesterberg et al. 2007, p. 16-22).

1.3 Principled Argument

This paper, as previously stated, has two purposes. The first purpose is to advance the use of visualisations as a means to make progress on faster more accessible pathways to statistical inference. The second purpose is to propose some forms of argument that we believe are critical for advancing statistics education and model their use in pursuing our first purpose.

Radical changes in statistics-education practice need to be debated widely in the open literature. On the big issues, questions of "Does it work?" in some phase-3-clinical-trials sense, are not meaningful. There is not even a consensus on what "to work" would look like, let alone having boiled down a myriad of possible implementations of different strategies to something simple enough to go into head-to-head comparisons using simple measures. For the issues we are thinking about, the field is

squarely in the phase 1 stage with occasional minor forays into phase 2 for detailed sub-problems. But these early stages are precisely those where widespread debate is most necessary. Useful components of such debates are principled argument and conceptual analysis.

By ***principled argument*** we simply mean setting down and then arguing from broad principles. Principles in statistics education are analogous to axioms in mathematics and assumptions in theoretically-derived statistical methodology. They facilitate critique and debate at two levels, the level of the principles themselves (cf. the applicability of assumptions) and at the level of implementation strategies (cf. correctness of derivations from assumptions). Debates that mix these things up lead to crossed wires and heat rather than light. It is more productive to move disagreements upstream to their source. Principles suggest ways of doing things (cf. corollaries). If principles can be overturned their corollaries will usually be of little interest. If they are accepted, or even tentatively entertained, we can argue about how well strategies and implementations embody principles. High-level debate based upon clearly stated principles is particularly needed if we are to identify "what we most value and why" and identify what we can dispense with and what we can streamline.

Some of the territory over which such argumentation should range is depicted in Fig. 1. An important factor that Fig. 1 omits (in the interests of not over-cluttering the diagram) is the role of practical experimentation and testing. Lessons learned from experimentation and testing can feed back into every node in the diagram. Experimentation can reveal assumptions that are not true, factors that have been overlooked and strategies that need to be changed.

Dearly held principles can, unfortunately, sometimes be contradictory, but this doesn't mean that one is false. We should be able to see clearly where contradictions occur to focus debate on how highly one principle should be prioritised with respect to another.

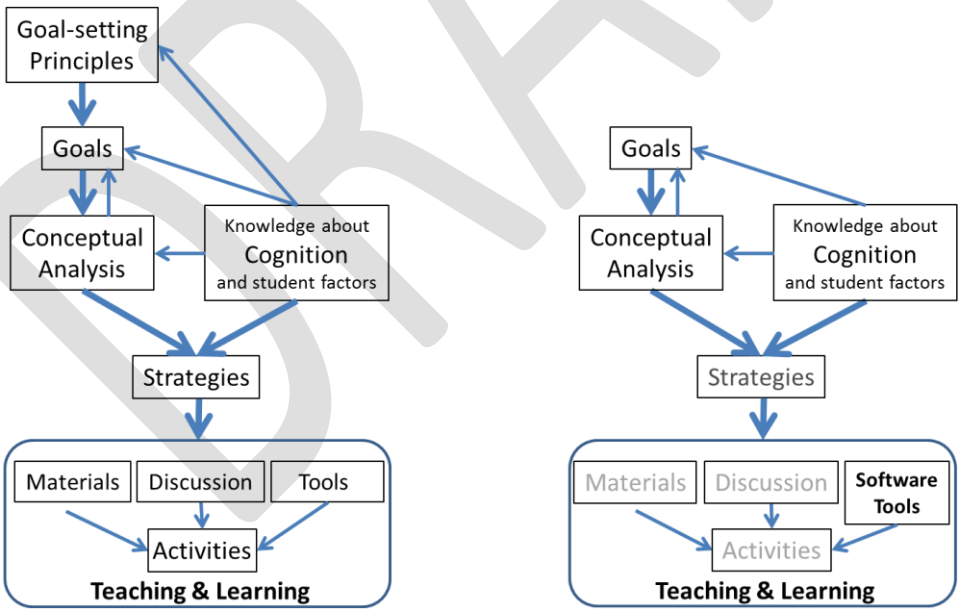


Figure 1. Factors for analysis, argument and design

2. PRINCIPLES

2.1 Goal-setting Principles

By goal-setting principles we mean meta-goals, bigger-picture goals that inform the goals for more detailed sub-areas. Our goals for the introduction of statistical inference by visual means are informed by a broader desire for statistics education that can educate and excite a very broad cross-section of students. Maximising accessibility is crucial to this so any goals have to be heavily informed by what we know about how people learn. We want to create an excitement about the potential of data and to increase the learner's capabilities for operating in the world of data; e.g. the capability to think, to do, to communicate and to learn more.

One educational principle that we believe to be well established is that "much is taught but little is retained". Learning is only really useful if has happened in such a way it will be self-activated when it is needed in future life (cf. Forster and Wild, 2010, Section 1.3; Pfannkuch et al., 2010, Section 1). Though we teach details we know they will be quickly forgotten. What we would most like to stick in the brain are broad principles and strategies, an ability to operate from them, and wider awarenesses that lead to inklings – those small voices that bubble up from the deeper recesses of memory saying, "I've seen something somewhere that might be useful here", then to our digging internally for what that thing might be, and then externally for more details if it seems promising (cf. Wild and Pfannkuch, 1999). The only realistic hopes from teaching details are primarily an increased facility with the types of thinking and argumentation used in developing them, and secondarily the hope that, having been worked through previously, they will be more quickly regained should the need arise.

If we want learning to be self-activated later in life we have to distinguish between what things we want to have burned indelibly into memory and what will suffice as vague inklings of a memory sufficient to set us on the path to locating a detail when we need it. Realistically the former can only be very small in number. But for them we need to search for ***abiding images***, which we characterise as mental images that still spring easily to mind long after the education that planted them, and conceptions that will initiate fruitful approaches by making them obvious – obvious what and obvious why. We characterise such conceptions as ***fertile conceptions***. ("How" is not absolutely needed for initiation. We can search for "how?" externally once we know "what".)

Another distinction that we think needs to be made is between the fundamentals of what statistics is and facilitating skills. The fundamentals, we would argue, concern the widely applicable big ideas of learning about the world using data. Mathematical argument and computing are supporting players – facilitating skills that can increase effectiveness – but not fundamentals of statistics. Where the time available is severely constrained, e.g. when students are likely to take only one semester of statistics, it is important to prioritise the fundamentals over facilitating skills.

For our intended audiences, our primary goal-setting principles are:

GSP 1: **Prioritise excitement about "what I can do with data and what data can do for me"**

GSP 2: **Prioritise increased capabilities for learning about the world using data and critique of data-based arguments**

GSP 3: **Prioritise the fundamentals of statistics above facilitating skills and use software that makes this possible**

GSP 4: **Prioritise big-picture conceptual understandings above details**

We will discuss goals with regard to introducing inferential ideas in Sections 3, 4.1, 4.2 and 5.

Of course goals and strategies in statistics education have also to be grounded in the realpolitik that balances what is desirable with what is achievable, the demands of clients and the way resources flow.

2.2 Conceptual analysis

Because of the severe limitations of any individual's working memory and because so little of what is "learned" will be retained for the long term we need analysis and debate about what we want those few things that we most want to stick to be. What are the few conceptual connections that matter most? Different ways of presenting and ordering a sequence of ideas draw in differing requirements in terms of foundational concepts and when they become required. This has a big bearing on when ideas and arguments become accessible. Conceptual analysis requires both theoretical argument and practical experimentation (see Section 1.3) which can reveal lurking foundational requirements that have been overlooked. The debates needed loosely resemble what goes on in mathematical proof. We lay down the pieces of an argument in a linear order in a way that exposes the bones of an argument so that it can be critiqued, dissented from, reargued by others. This also facilitates pinpointing where problems are occurring and planning workarounds.

The aspects of conceptual analysis (cf. Petocz and Newbery, 2010) we would like emphasised is analysis and debate about:

- the fundamental nature of concepts and how concepts inter-relate
- the concepts that need to be in place to support a desired new concept
- conceptual pathways (ways of sequencing the order in which students encounter concepts)
- prioritising concepts

The *action* principle for guiding teaching and learning strategies that we take from this is to seek pathways with fewer and easier steps.

CA 1: **Seek minimal conceptual pathways**

We will employ the elements of conceptual analysis above, in relation to introducing inferential ideas, in Sections 3-6.

2.3 General cognition

From the cognitive literature we draw particularly from the cognitive theory of multimedia learning (Mayer & Moreno, 2002) which is based on three learning theories: (1) dual coding theory (Clark & Paivio, 1991); cognitive load theory (Sweller, 1988); and (3) constructivist theory in which the learner is an active participant in making connections between prior knowledge and new experiences. The discussion to follow has almost entirely been extracted from Arnold et al. (2011) and we refer the reader there for further references.

Dual coding theory divides cognition into two systems, verbal and visual, and predicts that learning is enhanced when information is coded in both systems and connected between them. (There is some suggestion of another system related to other senses.) To process new information takes a great deal of attention. Both the attentional and working memory capacities of learners are severely limited making them major bottlenecks for cognition. Studies reported by Cowan (2000) suggest that the average person can only hold two to six pieces of information in their attention at once. Cognitive overload is an ever-present threat.

Three sources of *cognitive load* have been distinguished: intrinsic, extraneous, and germane. *Intrinsic load* is determined by the complexity of the application domain and by the learners' prior knowledge. The mitigating strategy is to increase preparatory knowledge. *Extraneous load* is the mental effort imposed by the way information is presented externally. We should strive to identify and minimise this. *Germane load* is the result of mental activities that are directly relevant to the learning. These activities contribute to learning and are relevant to the construction and automation of knowledge in long-term memory. Germane load is the most important factor in learning so the goal is to increase this. What is germane and what is extraneous is, of course, determined by an analysis based on the gaols of the learning experience.

Additionally, recent experiences are more accessible to working memory than time-distant experiences. The parts of an experience that require the most time and effort tend to dominate memories of, and take-home messages from, the experience. The metaphor of looking at a specimen under microscope is also useful. As we zoom in by increasing the magnification we lose sight of the natural environment within which that specimen lives and interacts. Focussing attention on any one thing crowds out our awareness and appreciation of others so making the details of anything understandable runs into the loss of the bigger picture as a law of unintended consequences. The *action* principles we have abstracted from these cognitive theories to inform teaching and learning strategies follow.

Cog 1: **Eliminate inessential sources of cognitive load and time & effort-consuming busywork**

Cog 2: **Minimise conceptual leaps and add complexity slowly**

To facilitate Cog 2 and aid overall sense-making:

a) Introduce things first in settings where as much as possible is familiar or obvious so attention can go to critical new elements
b) Build reminders of the most crucial enabling concepts into experiences that introduce something new
c) Hang new ideas off regularly revisited scaffolds of clusters of concepts
d) Iterate between zooming into a big picture for details and zooming back out so they assume their proper (often minor) place within the grander design
e) where there are weaknesses in arguments that students do not notice, postpone consideration of them until your main goals have been achieved

Cog 2a has implications for the use of language because introducing ideas expressed in terms of unfamiliar language that has itself to be learned contemporaneously places barriers in the path to understanding.

Cog 3: **Seek abiding images and fertile conceptions**

Cog 4: **Use active-learning strategies for the insights critical to long-term self-activation**

Both 3 (see Section 2.1) and 4 are targeting long-term retention and self-activation of` key essentials. Cog 4 advocates fostering learning experiences that lead to critical insights being realised (actively constructed by the learner) rather than passively received (reading or hearing).

2.4 Visual/graphical cognition

"When we are awake, with our eyes open, we have the impression that we see the world vividly, completely, and in detail. But this impression is dead wrong. … we apprehend only a tiny amount of the information in our surroundings, but it is usually just the right information to carry us through the task of the moment" (Ware, 2008, p. 1). Intake of visual information is very limited by the small buffers available for its storage. What we actually best see, it seems, are certain types of changes

against an expected background – unsurprising since such receptiveness is critical both for avoiding hazards and finding food. Changes against an expected background can be recognised by darting eye-movement scanning which seems to be at the heart of the visual system.

The following is summarised from the literature review of Arnold et al. (2011). For graphics comprehension there is a constant interaction between bottom-up perceptual processes of encoding information, which drives pattern building, and top-down inferring processes that are based on prior knowledge. Viewers re-inspect parts of graphics many times in the process of comprehension. Prior knowledge affects which parts of the graphic are fixated on and encoded, which then influences the inferences which are made. The more relevant prior knowledge students have before viewing an instructional graphic the better they will perform. The comprehension of graphics and learning outcomes are significantly affected by the display format. Visual cues such as colour should be used to highlight the relevant features to attend to, and fading should be used for irrelevant features. Motion attracts attention. In addition, Ware (2008) talks much more comprehensively about design strategies that make details in a display "*pop*" (jump into our attention). The additional action principles to inform visualisation-based teaching-and-learning strategies that we have extracted from research on visual perception, and cognition from graphics and animations (Arnold et al. 2011; *Educational Psychology Review* special Issue , volume 19, issue 3, 2007; Ware, 2008) are:

Vis 1: **Almost everything in a display should be as expected so that attention can be paid to the few novel features.**

a) Animations should only make changes in a small number of elements:

Vis 2: **Enable learners to identify and become familiar with relevant structures in representations before animation begins to allow the desired changes to pop**

a) Use of lead-in, hands-on activities can assist with comprehension and learning
b) Include controls in the animation itself that target lead-in familiarisation
c) Provide user control of the pacing

Vis 3: **Highlight features that particular attention should be paid to**

Facilitating strategies include:

a) using colour or bolding to highlight features that particular attention should be paid to
b) using movement (changes in size or position) to attract the eye
c) using fading to push "now-less-important", reminder elements into the visual background

Vis 4: **Dynamically move elements from one display into another to establish how the second display relates to the first**

Arnold et al. (2011) also reviews other important cognitive principles. For example, integration of information is most likely to occur if the learner has corresponding pictorial and verbal representations in the working memory at the same time. While issues related to verbal cognition and strategies for the integration of visual and verbal learning are absolutely critical they are beyond the scope of the current paper.

3 INTRODUCING STATISTICAL INFERENCE

We now begin some high-level conceptual analysis for the problem of introducing (classical) statistical inference paying particular attention to seeking minimal conceptual pathways (CA 1) and minimising conceptual leaps (Cog 2).

Cobb (2007) discusses the complexity of inference built on normal-distribution and normal-approximation-based sampling theory with its huge number of concepts all building on top of one another, its plethora of special cases and what Hesterberg (2006) describes as "its cookbook of formulas and assumptions". Additionally, "each adjustment we ask our students to learn takes their attention away from more basic ideas such as the fit between model and reality" (Cobb, 2007). The result is very slow progress and vast numbers of students who cannot see the wood for the trees and never really get it. Furthermore, "notice how little of any of this deals directly with the core ideas of inference! Randomized data production? That was chapters ago."

Cobb reserves particular opprobrium for the use of normal sampling-theory for the analysis of data from designed experiments involving, as designed experiments most often do, convenience samples. This, he suggests, is "fraud". Why is it fraud? Not because it is wrong. Normal theory is right in the sense that it often works. In statistics we are quite comfortable with using something because the answers it gives closely approximates something else (under appropriate conditions). Indeed Fisher originally justified *t*-tests for two samples by arguing that they would approximate the results of permutation tests (Erst, 2004, p.676). But it is fraud in the sense that it is incomprehensible without bringing to bear large numbers of concepts that have not been established in beginners. There is no direct conceptual connection between the random-assignment process used in data production and the random sampling theory used by the analysis. Our usual party line that with an experiment we are "inferring about the difference between what would happen if the whole population was given the treatment and what would have happened if they were all on control conditions" is of no help with convenience samples.

To understand that something works because it approximates something else: you need to understand what you should be doing, understand what the alternative does, and trust the evidence that the answers produced by the latter adequately approximate what you would get from the former – arguments involving spheres within spheres. Ptolemy would be proud. The aforementioned fraud matters because if we give up on comprehensibility all we are left with is incantation. Applied statistics is about trying to make sense of the world. We need to foster a disposition of ever striving to make sense of things and this is undercut when core tools being used emanate from the murky realms of dark magic.

Cobb (2007) calls for the logic of inference to be at the heart of the introductory course. "One of the most important things our students should take away from an introductory course is the habit of always asking, 'Where was the randomization, and what inferences does it support'?" Cobb (2007, p.3).

The above discussion suggests the following action principle for strategies for introducing statistical inference to the broad audiences we are targeting. As closely as possible,

SI 1: **the inferential method should mirror the process of data production**

In his discussion of Wild and Pfannkuch (1999), T.M.F. Smith (1999) worried about the tenuousness of random-variation models in situations beyond those in which randomness is induced by the study design. In models for observational data, in particular, randomness tends to enter as a modelling construct to help us cope with "unexplained variation" (Wild and Pfannkuch, 1999, Section 3 and p. 264). It is part of a what-if game. If the mechanism generating the data behaved like some proffered model containing random components, what would we then be able to conclude? This is very deep water involving conceptual complexity and no assurance that the real and model generating mechanisms are sufficiently similar in vital ways to give any assurance that the conclusions drawn from the model actually address the real problem. Statistical inference is on enormously more secure ground where it simply addresses randomness induced by the study design. It is also much easier to understand there. This leads us to conclude (Cog 1,2a) that teaching of inference should

deal first with actual random data-generation mechanisms – randomness by design – before addressing imaginary random data-generation mechanisms– randomness in models; the "is" before the "as if". We highlight this as an action principle for strategies for introducing statistical inference:

SI 2: **Address the "is" before the "as if"**

The ways in which randomness enters into data production that are most easily understood are random sampling to investigate populations and random assignment in experiments to facilitate causal inferences about the effects of interventions. These are also vitally important areas; the two areas we always emphasise most when talking about study design and data collection in introductory courses. Starting here also allows us to decouple exposure to the basic ideas of inference, which are difficult enough in themselves, from the additional complexities of modelling unknown data-generating processes statistically (Cog 1,2). The lessons about inference learned in a sampling context generalise fairly immediately to modelling contexts because in statistical modelling some model elements are treated as if they were sampled.

Applying action principle SI 1 above to these two situations leads to:

Data Production	**Base for inferential method**
*Random **allocation*** to treatment groups	*Random **re-allocation*** to treatment groups
*Random **sampling*** from population	*Random **re-sampling*** from the sample

With random assignment we can actually repeat the random operation used in data production. With random sampling we cannot and have to reach for something that seems to mimic it, at least in large samples. We go into this much more deeply in Section 4. A number of modern treatments also recommend using randomisation tests everywhere including for sampled data. But, in terms of accessible concepts, this is just the flip-side violation of SI 1 of using sampling-theory inferences for randomised experiments. There is no direct conceptual connection between the random sampling operation performed to get the data and a method of analysis based on random reassignment. (The procedure is justified theoretically as a conditional exact test; Ernst, 2004, p.681).

Randomisation tests and the bootstrap also bring another important advantage. Not only do they facilitate more direct connections between data production and inference, they also lend themselves particularly well to visual treatments of what they are, how they work and their operational characteristics – thus allowing us to stay far removed from the "formidable wall of mathematics" in the early stages.

These ideas are implemented in the VIT program (Section 1.1) in which we convey how bootstrap and randomisation inference works visually using the four modules which form the columns of Table 1. In parallel modules we display randomisation variation (random allocation, col 1) and randomisation tests (random re-allocation, col 2). Similarly, random sampling (col. 3) is paralleled by random re-sampling (col. 4, bootstrap). The quantities (statistics) being worked with cover and extend the range usually worked within elementary statistics. We have focussed on commonly encountered quantities that are relatively simple to motivate and understand, relate to features that can easily be seen or marked up on a "standard" plot of the data, and for which all of the desired behaviours can be conveyed visually. The beauty and strength of the bootstrap (respectively randomisation) is of a single, simple idea that works across multiple applications. It is therefore essential that, over and above conveying the nature of the inferential processes, the software should convey the "sameness" of what is happening across all of the listed situations to reveal the unity of the underlying way of thinking.

Table 1: Visual Inference Tools (VIT) modules

Track/investigate/use behaviour of …	VIT Module Name			
	Motivation of need for method **Randomisation Variation** *(**random assignment** of group labels)*	*Inferential **Method*** **Randomisation Tests** *(**random re-assignment** of group labels)*	*Motivation of need for method* **Sampling Variation** *(random **sampling** from Popn)*	*Inferential **Method*** **Bootstrap Conf. Ints** *(Random **re-sampling** from sample)*
1-variable				
Numeric			Whole Boxplot Mean, Median Quartile, IQR	Mean, Median Quartile, IQR
Categorical			Proportion (in selected level)	Proportion (in selected level)
2-variables				
Num | Cat (2 grps)	Diffs in: Means, Medians; 2-sample *t* -stats; IQR Ratio	Diffs in: Means, Medians; 2-sample *t* -stats; IQR Ratio	Whole boxplots Diffs in: Means, Medians; IQR Ratio	Diffs in: Means, Medians; IQR Ratio
Num | Cat (k grps)	Av. Deviation, *F* , "pseudo-*F* "	Av. Deviation, *F* , "pseudo-*F* "		
Cat | Cat (2 grps)	Diffs in: Proportions	Diffs in: Proportions	Diffs in: Proportions	Diffs in: Proportions
Cat | Cat (k grps)	Av. Deviation, Chisq	Av. Deviation, Chisq		
Num | Num	Regress. Slope Correlation Paired diffs	Regress. Slope Correlation Paired diffs	Regress. Slope Correlation Paired diffs	Regress. Slope Correlation Paired diffs

4. CONFIDENCE INTERVALS AND THE BOOTSTRAP

4.1 Motivation

We now begin dig down more specifically into ideas into some conceptual analysis of confidence intervals. Much of the discussion in this Section, although addressed to the reader, is really trying to lay down the shape of basic conceptions to be conveyed to beginners at a very high level. The biggest messages about confidence intervals are generic and have nothing to do with the details of a particular way of calculating them:

- Understanding why we need confidence intervals (or their Bayesian equivalents)
- Wanting to have them for every estimate we get or are given
- Knowing what properties they have
- Knowing what sources of error they do not correct for
- Being able to interpret a confidence interval in the context of a particular problem and communicate this well both orally and in writing
- Knowing that the bigger the sample taken the narrower the confidence interval obtained, (or the more precise is our estimation of the quantity of interest)

In elementary statistics, confidence intervals address the problem caused by sampling errors, the errors in estimation incurred where a population quantity is estimated using its sample counterpart. The *need* for confidence intervals is *motivated* by an appreciation of how (surprisingly) large sampling errors can be. The major problem to be solved is "my estimate is almost certainly wrong to some extent because of sampling error." The basic idea of confidence intervals as a solution to this problem is simple and intuitive. We allow for error by putting an interval around our estimate wide enough to accommodate the usual levels of the sampling error (usual in the sense of exceeded only in a fairly rare number of cases). This gives a range of plausible values for the unknown true value of the quantity of interest.

In real situations, however, the extent of sampling error is not something we can observe. We could still proceed in the same way if we could use a reasonable estimate of the likely extent of sampling

error obtained from information that we do have. Different ways of constructing confidence intervals employ different ways of solving the *sub-problem* of estimating the extent of sampling error. One widely applicable way is via bootstrap re-sampling. Other ways emphasised in elementary statistics are based on mathematical theory deduced from assumptions like sampling from a normal distribution.

The ***visual*** *motivation* for using the *bootstrap* to *solve the* "unknown-extent-of-sampling-error" *sub-problem* is that when we sample from known populations (which is where we can actually observe the extent of sampling variation) the patterns of variation for an estimate we see when sampling from the population look very similar to the patterns of variation we generated by re-sampling from the sample (see Fig. 2). More snappily, in the cases where we can see both, "re-sampling error looks a whole lot like sampling error". This suggests that we use *re-sampling error (which we can see)* to *estimate* the extent of *sampling error* (*which*, in real applications, *we cannot see*). Why would anyone have thought of looking at re-sampling the sample in the first place? Because the distribution of values in the sample is the best approximation we have to the population distribution. Fortunately, sampling variation and re-sampling variation patterns can be quite similar even when plots of the population distribution and the sample data look rather different.

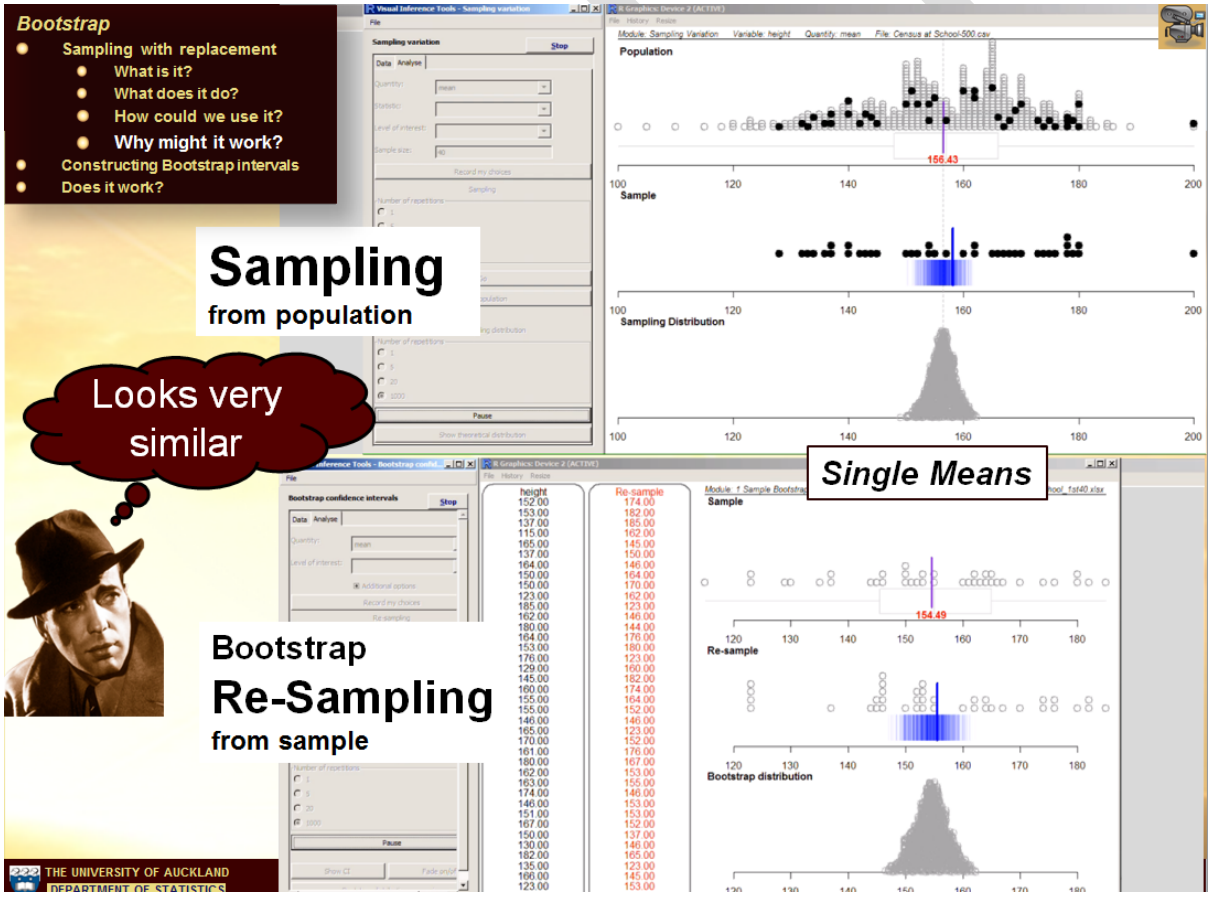


Figure 2. Sampling variation and bootstrap re-sampling variation in sample means
(Animated gif at http://xxx)

With bootstrap re-sampling there is an immediate and therefore understandable relationship between the sub-problem we have (needing an estimate of the unknown extent of sampling error) and the proffered solution (estimate it using re-sampling error). That relationship is simply, "they look very similar."

It is time now to step back and qualify some of the assertions made above. For these ideas to work, students need to understand what sampling error is and understand displays of the patterns of variation it produces (Section 4.2). They need also to understand what bootstrap re-sampling is, ideally have some idea of why it could conceivably behave somewhat like sampling, and be able to understand displays of the patterns of variation it produces (Section 4.3). This is deceptively difficult. Most students do not, for example, understand the usual displays of sampling variation as sampling distributions, nor how they relate to a primary display of their data (see Wild et al., 2011).

What we have described above is simply *motivation* for bootstrap confidence intervals, *not justification*. It suggests a possible "solution" but confers no assurance that it will work. This begs the follow-up question, ***"Does it work?"*** to which our response is, ***"Simulate and see."*** When we simulate we often find that the techniques still work reasonably well beyond those situations where the patterns of variation look very similar but that is not part of the equation, it is just a bonus. As a pedagogical aside, the visual motivation does not work well if the sample sizes being used are too small, or if we are using quantile-based estimates with data that has been too heavily rounded.

The situation we face with the purely visual approach outlined above is not entirely dissimilar from what happens in "grown-up statistics". The theory behind most methods, and particularly those that do not rely on strong distributional assumptions, is asymptotic (as sample sizes tend to infinity). As such, the theory merely suggests something that might work in practice. Performance with finite samples has to be investigated by simulation. So-called exact theory involving strong distributional assumptions also merely suggests something that might work in practice. Sensitivity to the inevitable violations of these assumptions has also to be investigated by simulation.

We have been asked, "But isn't re-sampling confusing? Sampling and re-sampling variation looks similar. Won't students get confused when we show them both?" This misses the main point. The similarity between these patterns is why bootstrapping makes sense and why it works. The relationship between these two operations is an absolute fundamental that needs to be a focus of teaching reinforced by assessment. We want to allow for the extent of sampling error which, unfortunately, we cannot see. We use the extent of re-sampling error, which we can see, in its place. Simulation shows this generally works very well.

4.2 Strategies for visualising sampling variation

We now specialise to consider graphic strategies for showing how estimates are affected by repeatedly sampling from the population, leading to design decisions used in VIT's sampling variation module. As can be seen in Fig. 2, most of the graphic elements used to convey sampling variation is re-used identically in conveying the effects of re-sampling from the sample in the bootstrap (Section 4.3).

We are interested in sampling error viewed as variation in estimates about the true (population) value. A fertile conception here would be provided by a mental image that would make "I need an uncertainty band around this" shout out at you whenever you saw estimate, long after the learning experiences that transmitted it. Conventional displays of sampling distributions of statistics or estimates do not do this. Frenetic movement revealing a random process unfolding over time and leaving behind its history – a smudge of accumulated "footprints" on primary graphical displays of data as in Wild et al. (2011) – has shown promise in preliminary research (Arnold et al., 2011; Pfannkuch et al., 2013) for acting like this and we extend and make heavy use of this idea here. Because the conventional sampling variation displays are still very useful and are more widely applicable, however, we want to add those to the students' repertoire as well.

The connections we are trying to establish are

- What sampling looks like in the setting we are investigating
- How variation in samples translates into variation in estimates
- Ways of seeing the patterns of variation so generated, including seeing these patterns in terms of discrepancies between the estimates and the target true value

As in Wild et al. (2011) we sample from real populations not abstract theoretical distributions (Cog 1,2). Arnold et al. (2011) introduce sampling in a hands-on way using population bags containing multi-variable data cards, one card for each individual in the population. When our goal is inference, however, we need to be able to see sampling not only in physical terms but also in connection with "distribution", something which the image of sampling cards from the bag does not provide. We also want to do this intuitively without any formalisation of "distribution". For continuous variables this can be done by using stacked dot plots starting with finite populations small enough that you can represent and see all population members individually (e.g. Fig. 3, graphics panel 1). With binary data the same ends can be realised using segmented bars "populated" by representations of every individual in the population (cf. the top graphics panel of Fig. 4(b)). In the hands-on world, plots constructed from the data cards themselves can be used to establish the connection between points on a plot and the "person behind the data point".

Fig. 3 comes from the end stages of an animation of sampling variation in a mean from VIT's sampling-variation module. Although means are used here everything works in exactly the same way with medians or quartiles. Animations are impossible to describe adequately with static media, so this paper supplements its verbal descriptions of animations by linking to dynamic content. A narrated movie showing and explaining the animation that Fig. 3 comes from is given at http://xxx. The movies of this and other animations also go much more deeply into what is happening and why than is possible in the written paper.

Panel 1 of Fig. 3 (Population) displays the heights from a school of 500 students using a stacked dot plot. This set of students is being used as a population to sample from so what we are seeing is a population distribution of heights. When the animation starts all of the graphics panels are blank except for panel 1. In it you can see every individual in that population represented by an open circle and arranged in height order. The position of the population mean is marked with a heavy vertical bar and also a vertical dotted line that goes down through all panels. As individuals are sampled their appearance changes (to a black disc), visually connecting the act of sampling and population distribution. When sampling is complete the set of sampled individuals move down into panel 2 (Sample) which now contains "our data" (Vis 4) and the position of the sample mean for that data set is marked. As further samples are added "footprints" of all means previously seen accumulate below the dots of the current plot. To cater for conversations around the fact that we do not normally see what is happening in the sampling process the "fade population" button makes what is happening in panel 1 only dimly visible as if viewed through a veil.

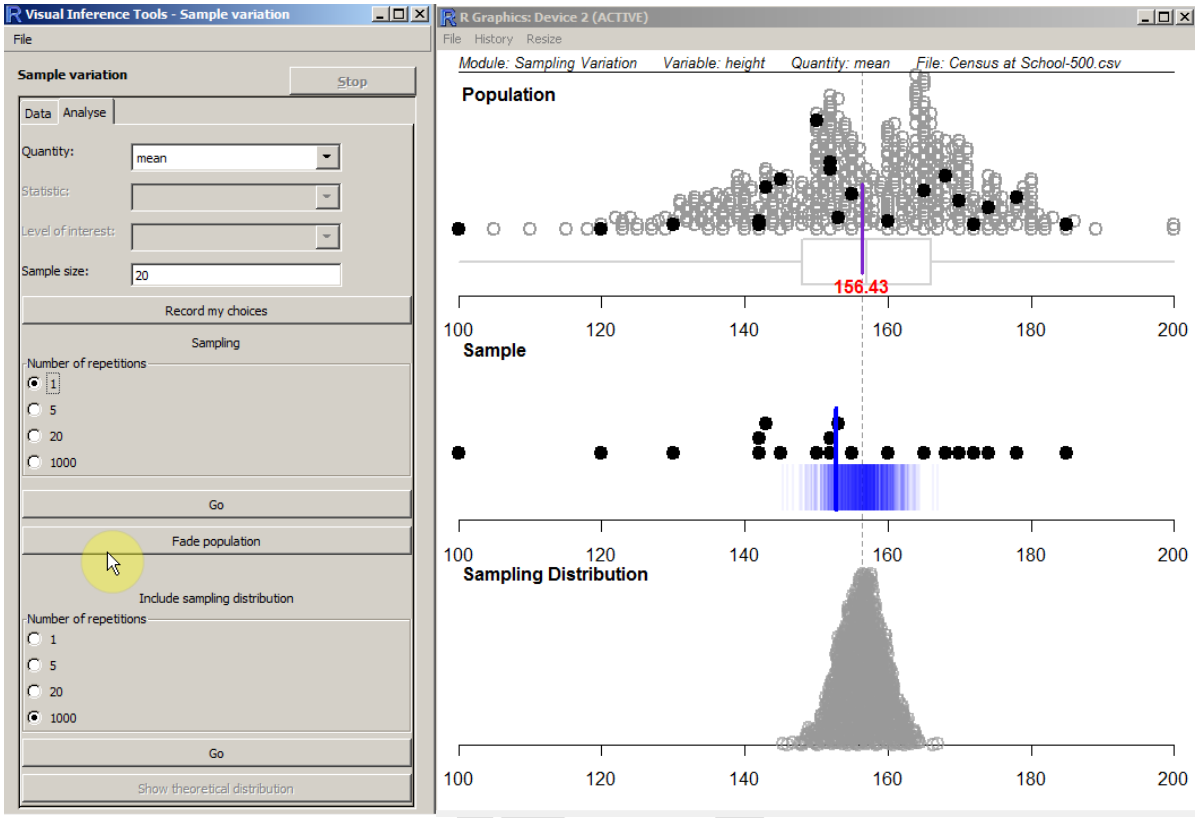


Figure 3. Sampling variation in a mean
(Narrated movie of animation at http://xxx)

We need to distinguish between understanding *the nature* of the operation being performed, namely, taking a sample and plotting the position of the mean and understanding *the effect* of sampling on the positions of those sample means. When 1 or 5 repetitions are selected in control panel 2 all elements of sampling, dropping down, position-marking and footprint-accumulation are shown sequentially in detail to cement in an appreciation of the nature of the operations being performed (Vis 1,2). When 20 or 1000 repetitions are selected, sampling and plotting happens instantaneously so attention can switch from the nature of the operation to the effect of the operation. The frenetic, erratic movement of the sample means gives an (abiding?) image of the nature of the randomness of the variation and the building up of the footprint "smudge" shows the extent and relative densities in the variation.

For many purposes panel 2 (Sample) conveys enough without going further but this module also deals with more conventional displays which are formed in panel 3 (Sampling Distribution). Similarly to what happens in panel 2, if 1 or 5 repetitions are selected in control-panel 3 the animation in display-panel 3 is designed to emphasize understanding the nature of the way the display is built up. Mean-bars drop down from panel 2 into panel 3 transforming into points as they do so and begin to form a stacked dot plot (Vis 4). When larger numbers of repetitions are selected the dropping down is turned off so that the focus can switch to watching the building up of the sampling distribution in panel 3. All three panels are in action at once giving reminders of all of the earlier connections including that between population and sample plots and the random movement of the sample means.

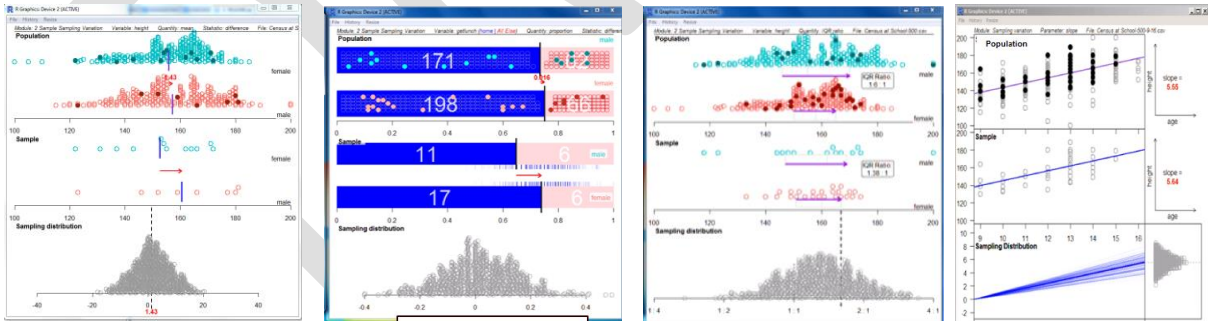
Sampling variation is being displayed in two different but equivalent ways. In panel 2, the density of accumulated sample means is conveyed by depth of colour (from overprinting of semi-transparent bars) whereas in panel 3 it is conveyed by the height of the stacking. The display in panel 2 has the advantage of being directly connected to the primary data plots students look at and the frenetic

jumping about seen as new means appear conveys random-appearing variation, and therefore uncertainty, in the same visual space in which we view the data. All of this is seen in close juxtaposition with the population information in panel 1.

While the panel-2 display has the advantages of immediacy and "being in the same visual space as the data", it is an idea that only applies to quantities (like means and quantiles) that can be represented as marks against the scale of the primary plot. There are other features that can be clearly seen on the primary plot but cannot be represented simply by marks against the scale. Examples include interquartile ranges, which are clearly visible if we superimpose a box plot on the dot plot, and differences between centres in two-sample comparisons. Ratios of interquartile ranges, as a way of considering comparisons of spreads, can be made visible with simple mark ups. Slopes in regression are clearly visible.

For quantities that, although clearly visible, cannot be represented as marks against a scale we have to move to more abstract representations of sampling variation. It makes sense to build up an understanding of these representations in situations where we can see the simpler representations as well, and also to carry over as much as possible from the visual environment used in the simple cases to the more complicated cases so attention can be confined to those elements that are intrinsically different. Distances between centres or proportions are represented as directed arrows which drop down to form the sampling distribution. Ratios of box widths in boxplots can "almost be seen" in that we can have a rough idea by looking of how much longer one box is than another but this then has to be converted into number (the ratio) and it is these numeric ratios that drop down from panel 2 into panel 3 to form the sampling distribution. Vignettes from sampling distribution animations for some other data features are shown in Fig. 4. To see them at work click to the links to the accompanying narrated movies. An additional movie for a single proportion is given at http://xxx. It would be helpful to view the single proportion before looking at differences in proportions http://xxx.

All of the graphic strategies described here for the sampling-variation module are used essentially identically in the formation and plotting of bootstrap re-sampling distributions in the bootstrap confidence interval module, randomisation distributions in the randomisation-variation module, and re-randomisation distributions in the randomisation-test module (Vis 1).



(a) (b) (c) (d)

Figure 4. Vignettes from some other examples of sampling variation: (a) differences in means, (b) differences in proportions, (c) Ratios of IQRs, (d) Regression slopes (Movies of corresponding animations are at http://xxx, http://xxx, http://xxx and http://xxx)

4.3 General patterns across the visualisation modules and their rationale

We make a very brief digression to capitalise on what has been learned during the discussion of VIT's sampling variation module to describe unifying patterns applied across the whole set of modules. This unification is important in setting animations in a visual environment in which as much as

possible is as-expected (Vis 1), thus avoiding unnecessary distractions and allowing attention to be focussed on the few essential differences. As illustrated in Fig. 5, we repeatedly use a three output graphical *display panels* operated using corresponding *control panels* on the left-hand side. Data is read in by the user and displayed in panel 1. For example, in the sampling variation module it will be treated as population data to be sampled from, whereas in the bootstrap module it is data from a sample. Control-panel 1 caters for choice of settings such as quantities whose behaviour is to be investigated.

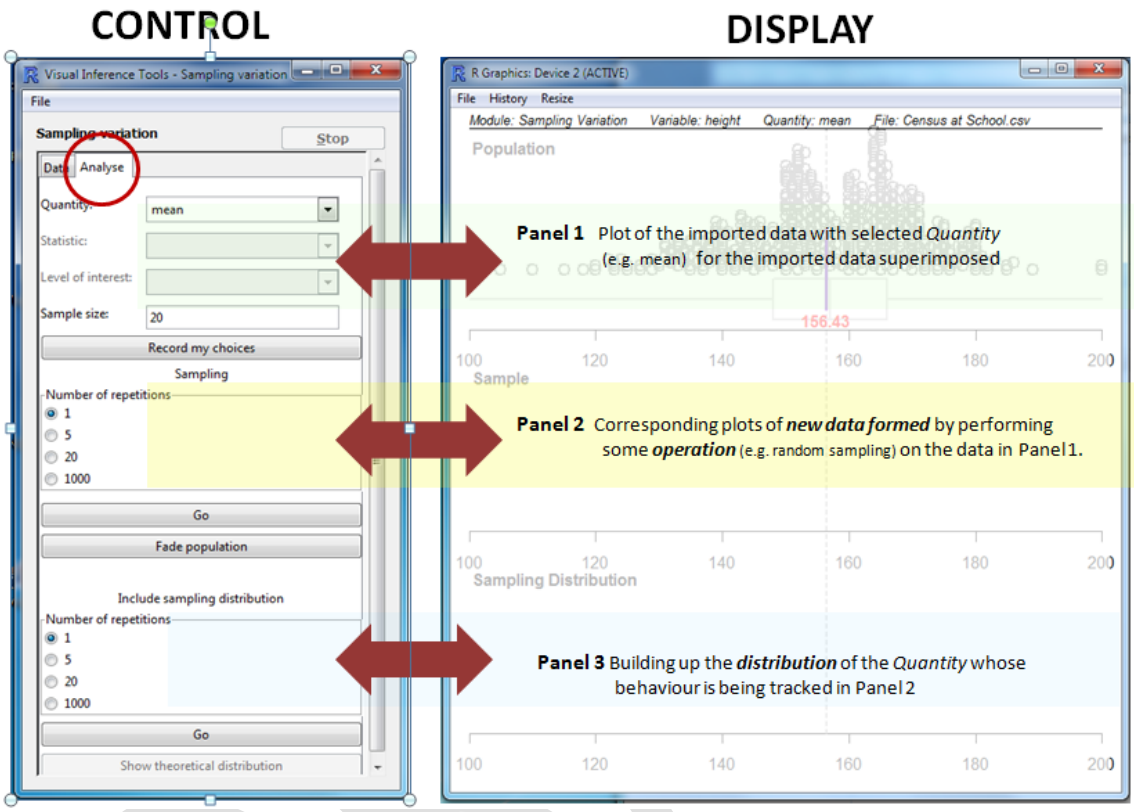


Figure 5. General layout design of VIT modules

Display-panel 2 displays appropriately what is produced by some operation applied to the input data. In the sampling-variation module the operation is based on random sampling; in the bootstrap module the operation is resampling with replacement; in the randomisation-variation module it is the random allocation of group labels; whereas in the randomisation-test module it is random re-allocation of group labels. A key feature of interest is highlighted by mark-up in both Display-panels 1 and 2 (be it a raw summary measure, or a differences, or a ratio, …). These operations, and where and why they are applied, are statistical fundamentals which should be at the core of the learning processes VIT is being used to facilitate. Control-panel 2 gives control over the number of repetitions to be generated in display-panel 2. When a small number of repetitions is chosen, animation is at a very detailed level because the intended focus is on understanding *the nature* of the operation being performed. When the number of repetitions chosen is larger very detailed-level animation is suppressed and playback is sped up because the intended focus is on understanding *the effects* of the operation being performed on the key feature of interest.

Display-panel 3 captures the *behaviour* of the key feature of interest captured in display-panel 2 as a *sampling distribution*. This panel is controlled by control-panel 3. When small numbers of repetitions are selected the focus is on how the target feature of the panel-2 plot turns into a point in the panel-3 sampling distribution plot. When larger numbers of repetitions are selected detailed-level

animation is turned off to focus on watching the building up of the sampling distribution in panel 3. All three panels are in action at once giving reminders of all of the earlier connections.

Depending on the module additional information such as confidence intervals or tail proportions becomes available for mark-up display once animation is complete. Theoretical curves can sometimes be added to distributions in panel 3 to facilitate connections with theoretical approaches to the corresponding problems.

Slight modifications to the Fig. 5 layout are made in the confidence-interval coverage module in which panel 3 captures coverage history rather than a sampling distribution, and applications to regression slopes where the standard panel-3 position is used to accumulate the slopes and the distribution accumulates off to the right with a visual connection to the slopes captured (see Fig. 4d).

4.4 Visualising the bootstrap

Even if they have heard the sampling-with-replacement language, students do not come with a pre-existing appreciation of what (re-)sampling from the sample with replacement is, how it acts and how it produces variation in the bootstrap estimates. Such an appreciation is a prerequisite both for visually motivating the bootstrap as in Section 4.1 and for understanding how confidence intervals are generated from bootstrap resampling.

The component of bootstrap resampling that is universal is resampling the units (e.g. people) in the sample. This corresponds to resampling the rows of the data set (for a standard cases-by-variables representation of data). In stark contrast, the ways graphic elements move from the sample to the resample differ greatly from situation to situation. So by default the parts of the VIT animations intended to establish the nature of the bootstrapping operation emphasize the selecting and then moving of data-set rows in to place to form the resampled data set. This occurs in two additional panels added to the left-hand side of the display window in Fig. 6. The first data-panel contains the actual data. The second panel labelled “Re-sample” is the destination panel that data-set rows selected in the re-sampling process then move in to. The graphs in panel 2 are then, by default, just a result of plotting the resampled data set. The ability to track how points move is an optional extra for occasional use. Another option steps through showing where each data point in the sample went to in the resample, thus enabling a teacher to further underline the nature of sampling with replacement. Some data points are not used, some others appear multiple times (see http://xxx).

Apart from the details of how resampled data is constructed, the remainder of the animations in the bootstrapping module play out identically to their counterparts in the sampling variation module. The only difference between bootstrap and sampling-variation animations corresponds to a fundamental conceptual distinction at the heart of the required learning.

Additional information becomes available when the bootstrap distribution has been formed. At that point we can ask for the formulation of a percentile confidence interval. The tails of the distribution are trimmed off and the interval formed by the panel-3 axis. Whenever possible, the intervals are moved up onto the data in panel 1 itself, putting the interval back into the visual context of the data. This can easily be done where the quantity being estimated can be represented by a mark against the scale of panel-1 graph. Otherwise there are problems, such as the lack of a natural origin for representing the interval for a difference on a two-sample plot.

Once the interval has been added, we give the capability to push all of the information relating the constructing the bootstrap interval into the visual background by fading it out so that just the data, and the point estimate and the interval around the point estimate are clearly visible (Vis 3c). Everything else is analogous to the working in long division. It is an aid in getting the answer but

once we have it we put all that working behind us. It is only the answer that we take forward and use.

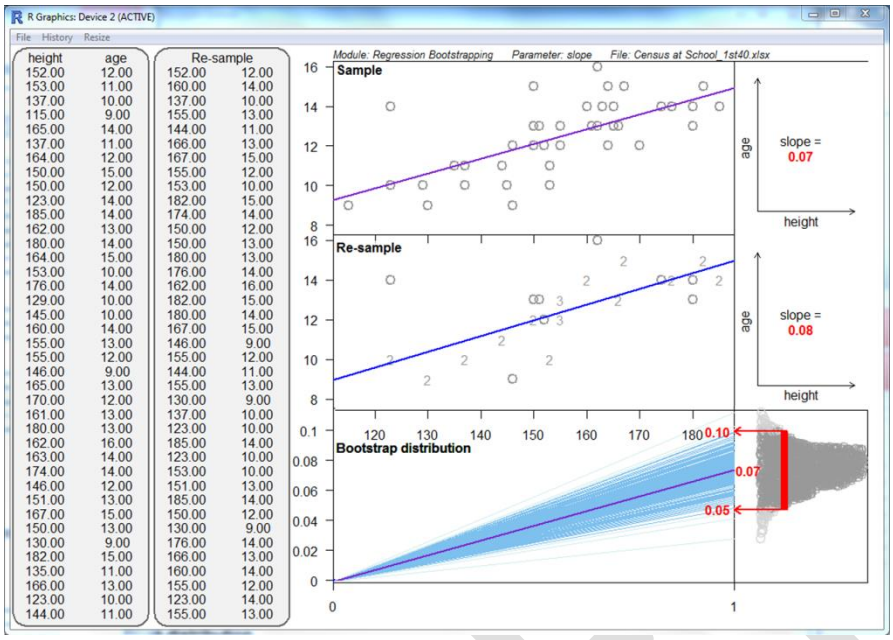


Figure 6. From bootstrap confidence interval formation for a regression slope (prior to fading) (Narrated animations: http://xxx emphasising resampling, http://xxx emphasising CI formation)

We have already discussed the advantages of the sameness of bootstrap and sampling-variation visualisations, i.e. maximising commonalities between modules. The beauty and power of the simple bootstrap comes from it using the same, simple idea every time. Underlining this point necessitates maximising commonalities within bootstrap visualisations across applications. By showing the bootstrap idea at work in a whole lot of situations that are understandable (because you can actually see everything that is going on) we prepare the ground for the leap to "this is a general tool I can use fairly routinely (without any need for visualisation)". In practical terms, the main purpose of the visualisation is to foster a sense of familiarity and comfort with the technique and its unified way of thinking, prior to routinely accepting the results it delivers. It is something that makes sense. It is not just magic incantation.

Regarding, "Does it work?", or more properly "When does it work?", we give the ability to experience the way investigating coverage via simulation works. Beyond that point it is not unreasonable to be asked to accept what others have found about where things do and do not work – understanding how these things can be checked and knowing that the ability to check them is there if we so desire. After all, this is how it works in "grown-up statistics."

The 1,000-replications buttons on the visualisations also provide an analysis tool for early experiences using these techniques, a tool that makes it possible to see not just "the answer" but also to probe into how it was constructed, if so desired.

4.5 What we do *not* need to know

In this development we do not need to know (CA 1, Cog 1,2) anything about formal ideas of distribution. They are finessed away by the nature of the graphics. We don't need abstractions like

histograms which are a further step removed from the raw data than are stacked dot plots in which you can see the raw data points. More valuably we do not need to understand anything about theoretical distributions, several abstractions too far for many students. We do not need to know anything about the jargon of parameters and estimates. We can replace this with much more direct language that explicitly addresses the main issue, namely what is happening in the sample versus what is happening in the population. Continual explicit repetition of the population/sample language (e.g. sample-mean and population-mean) in the early stages continually reinforces what we are doing whereas "parameter" obscures it. The development here has many fewer concepts to link than theory-based approaches. We have tried, moreover, to strip it down to its absolute barest essentials. And yet it can give experience with a wide range of important applications almost immediately. Issues of sample size as it relates to nominal and actual coverage frequencies cannot be circumvented, however, but are in common to both bootstrap and standard theoretical approaches.

With the percentile bootstrap we do not even have to know what a standard deviation is! The percentile bootstrap has the primary advantage for our purposes of corresponding to information that is immediately available visually, unlike something based on standard deviations. (A secondary advantage as automatic adaptation to asymmetry.) But after the idea of standard deviation and its connotations have been put in place then bootstrap standard errors become accessible as do confidence intervals based on plus or minus *K* (bootstrap) standard errors. Since bootstrap standard errors for means and proportions are essentially the same as the usual standard errors (to within $\sqrt{(n-1)/n}$) this gives us a valuable bridge between bootstrap and conventional normal-theory methods.

The more switches and choices you add to software the more difficult the software environment becomes to understand and use. In our teaching we postpone any consideration of levels of confidence until after the levels of exposure described here and the software does not cater for them.

5. RANDOMISATION TESTS AND DESIGNED EXPERIMENTS

5.1 Motivating and visualising random allocation

We now begin some conceptual analysis for randomisation tests. The discussion here envisages a student audience at a point where they have already learned some of the basic ideas about randomised experiments and in particular that, in the presence of unknown factors giving rise to differences between individuals, they are the most reliable way we know of making fair comparisons and establishing causal effects due to treatment differences. In this section we confine attention to the simple completely randomised design.

Our jumping off point is the fact that while randomised experiments may be the best way we know of making fair comparisons, they are not perfect. If we simply randomly allocate group labels to our units to form artificial "groups" we can get quite large apparent differences between the "groups", even though there are no real differences at all. All the differences we see are just due to the luck of the draw as to who ends up being placed in what group. This is the primary motivation for significance testing with experimental data. If random labelling alone will produce differences bigger than the ones we see in our data reasonably often then it makes no sense to claim that our data (alone) tells us anything about the existence or direction of treatment differences.

The randomisation-variation module of VIT has been designed to enable students to experience randomisation variation as a random process unfolding in time and to see the sizes of differences

that random allocation of group labels alone (i.e., and nothing else) will produce. This is done to provide motivation for why we need something like a significance test in general, and to motivate the randomisation test in particular. As Cobb (2007) says in his abstract, "Randomization-based inference makes a direct connection between data production and the logic of inference that deserves to be at the core of every introductory course."

Fig. 7 comes from taking the first 100 students in the school data and artificially randomly labelling half of the students as belonging to group A and the rest as group B. We are looking at differences in the proportions of each group who brought their lunches from home. The animations work as described in Section 4.2 to establish the nature of the operation (randomly applying group labels), its effect on differences in group proportions seen dynamically on the primary graphic we are using for this type of the data, and capturing the difference arrows in panel 2 to build up the randomisation distribution in panel 3. Panel 3's differences axis has the same scale as the data axes in panels 1 and 2 but is centred at zero. We are seeing some quite big differences with the bulk falling between ±0.2 (±20 percentage points).

Fig. 8 comes from the randomisation-test module near the end of the building up the randomisation distribution. The application depicted is, essentially, a 1-way analysis of variance on a continuous response with 3 groups. The essential difference between the motivational randomisation-variation module and the randomisation-test module is that in randomisation-variation there is no actual treatment-group and therefore no group membership distinctions made in the panel 1 plot, whereas in the randomisation-test output we see the actual group-membership values displayed as part of the data. Group-membership distinctions are now a critical part of the panel-1 plot because what we want to do is compare actual group differences in panel 1 to those generated by random relabeling in panels 2 and 3. Fig. 9 shows some of the changes made to the panel-3 plot (conflated over several frames) as we make this comparison. The summary measure of intergroup differences from the data is at about the upper 1% point of what we get from random re-labelling.

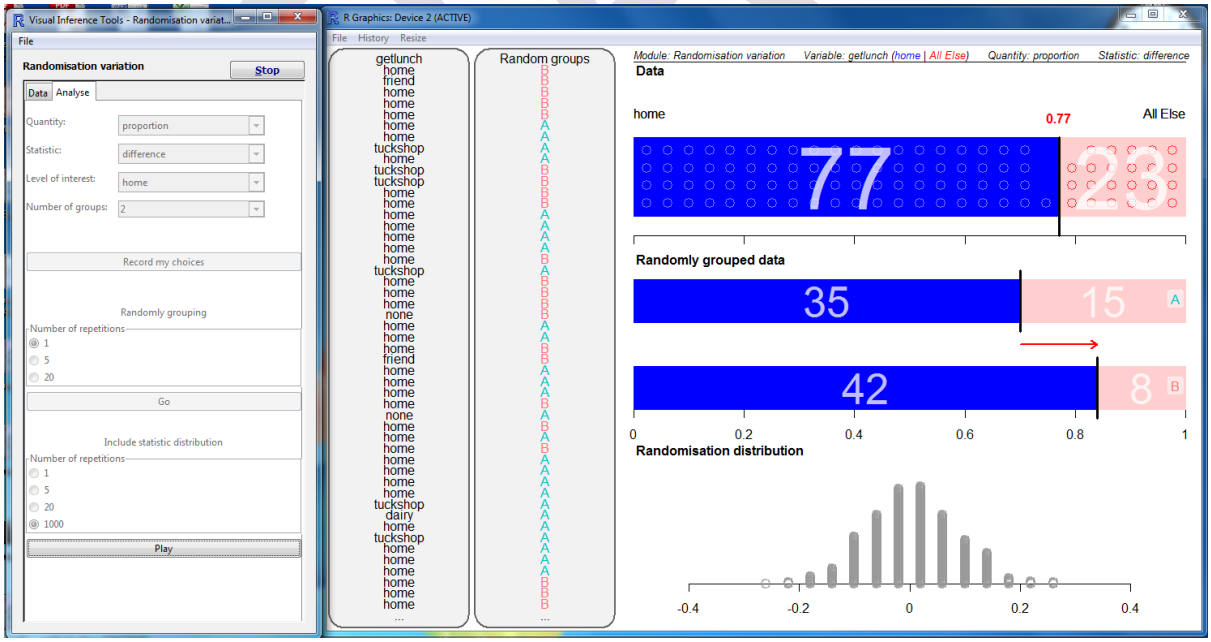


Figure 7. Randomisation variation for differences in proportions (artificial random grouping) (Animated gif at http://xxx)

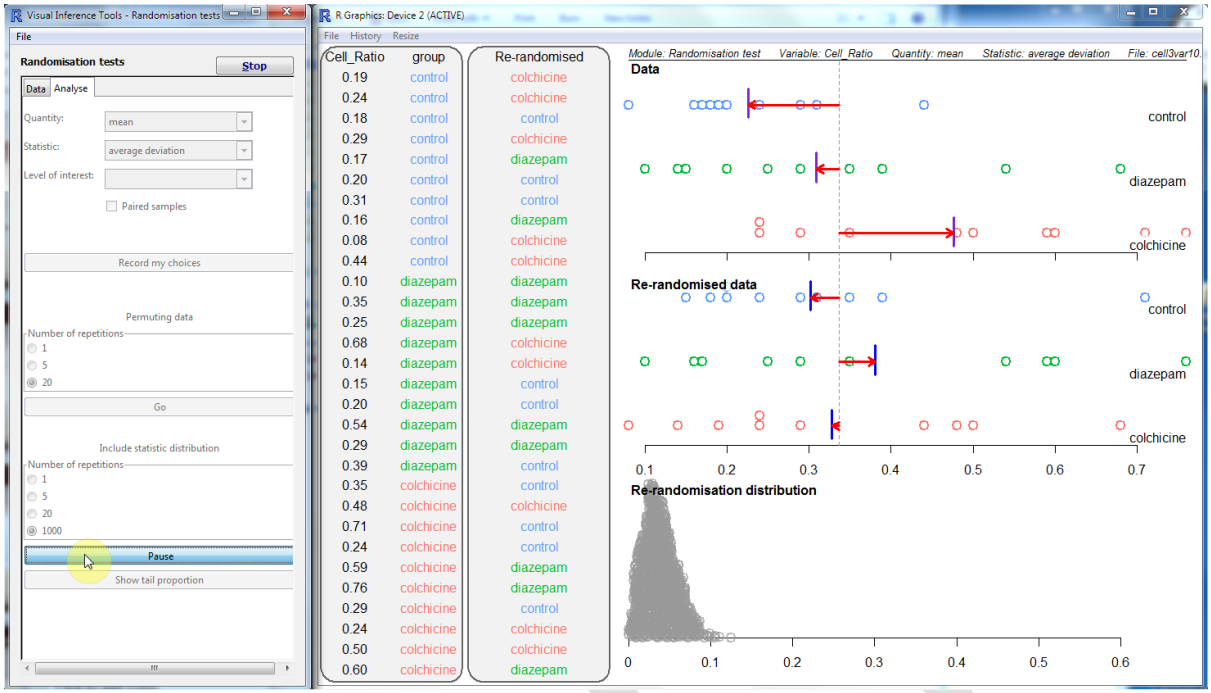


Figure 8. Randomisation test for differences between centres of 3 groups near the end of animation (Animated gif at http://xxx)

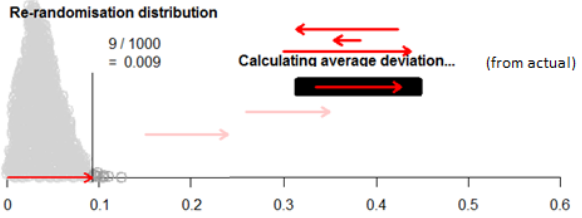


Figure 9. Information added over several frames to panel-3 plot of Fig. 8 after "Show tail proportion"

We saw that bootstrapping has a single, simple, defining invariant to be emphasised in the animations, namely the random re-sampling of the rows of the data set to form the bootstrap data set. Randomisation tests likewise have a single, simple, defining invariant – redefine the groups by replacing the actual set of group labels by a *new set* of group labels obtained by randomly re-allocating the original labels. This should, therefore, be a central unifying characteristic to be emphasized in randomisation-test visualisations (see the data-display aspect of Fig. 8) so that they can convey the unified underlying way of thinking. To see how this is done in VIT's randomisation-test module, see http://xxx . Once again the 1,000-replications buttons on the visualisations also provide an analysis tool for early experiences, a tool that makes it possible to see not just "the answer" but also to probe into how it was constructed if so desired.

5.2 Rationale and details

With a (completely) randomised experiment, we have randomised our experimental units into groups and then we have done something different to each group (applied the treatments). The different treatment we have applied to each group may have acted to make the groups different in systematic ways or may not have produced any differences at all (either because the treatments are either totally ineffective or because they act identically). Superimposed on top of anything the

treatments have done in terms of creating real group differences are random perturbations caused by random allocation, the sort of bouncing around we see in the randomisation-variation module.

Is the effect of random allocation acting alone ("chance alone") sufficient to produce the size of group differences we are seeing in our data? Or are the observed group differences sufficiently large to convince us that something else is going on as well? If differences of this size or even bigger would quite often be produced under random-allocation acting alone we have no evidence of real differences being produced by the different treatments. If they are larger than what is produced by random labelling then, unless there has been a breakdown in the running of the experiment, the only other thing that can be acting to make the groups different is the treatments we have applied. All other differences we might see between the groups can only be there by the luck of the random-allocation draw and thus form part of the chance-alone perturbations.

There is an essential "suppose" or "what if" element to the randomisation test. "Could we see differences this big if only random allocation was happening and *nothing else*?" Now "*nothing else* is happening" means that the treatments are making no differences, or equivalently no matter what treatment was given the result would have been the same, so the effect of random allocation could be investigated by using the data we have and randomising the group labels, just as in the VIT visualisations. The test is performed by looking at where a group-differences measure for the data sits with respect to the distribution of such measures generated by random re-allocation.

The above are what we see as the high-level, biggest ideas involved in randomisation tests. One core element is a measure of "group difference(s)". With a continuous response and only two groups, prior experience will already have taught students to think in terms of a difference in centres (using either mean or median). This works well for visualisations because they are features that are easily marked up on standard plots and their positions make visual sense (the mean as balance point, or the median as half-way point in terms of number of observations). We use a directed arrow to convey the difference between, or shift in, centres. The same thing works for a difference in proportions. Plots may also stimulate a desire to look at a change in spreads. For this we use a ratio of interquartile ranges and the graphs essentially use a log scale (e.g. … , 1:3, 1:2, 1:1, 2:1, 3:1, … equally spaced).

Things get more difficult with more than two groups. We like to start as much as possible with things that you can see and use to mark up a plot (Cog 2a). Ideally the distance measures to be used would be so obvious that we would not have to think or talk much about them. In that vein our default combined-difference-measure for multiple groups is the average length of the set of group-deviation arrows that you can see in Figs. 7 and 8. This is also done for binary responses (multi-group differences in proportions). Essentially we are trying to take "complex discrepancy measure" out of the conceptual pathway as a potential impediment to initial understanding. Once the testing idea is established using this simple, visual, discrepancy measure, moving to the conventional *F*- and chi-square discrepancy measures adds nothing conceptually new to the testing procedure. Motivating why it is better to use such complicated discrepancy measures instead of simpler ones is a challenge, of course, but one that applies only to an optional detail and not to the big picture. Having it correspond simply to a choice in a drop down menu underlines this.

No matter what choices are made, the difference(s) we see in the data in panel 1 are dynamically transformed into a position of the scale of the randomisation distribution and attention is drawn to "how do they compare?" by first the dynamic moving of them down into panel 3 and second by colouring in the relevant tail of the randomisation distribution (cf. Fig. 9). This is a "closeness-to-the-edge" measure communicated both visually and numerically as the observed proportion of times random re-allocation produced something at least this extreme (inclusive "or"). How the conversation proceeds from there is up to the teacher. In the two extreme cases of either, "randomisation (almost) always produced something smaller than this", or, "randomisation often

produced something bigger than this", the inferences to be drawn are fairly obvious. When something less clear cut crops up the teacher can lead the conversation towards the desirability of formulating some sort of operational rule based on the tail areas.

5.3 What we do *not* need to know

As in Section 4.5 the beauty of what is presented here resides partly in the number of important applications we can deal within a simple, uniform, visual setting. But more importantly it resides in "what we do not need to know" – or at least until after a great deal of experience with the basics of testing has been gained. Again we do not need to know anything about theoretical distributions and the jargon of parameters and estimates. We do not need the abstract notion of a null hypothesis. We can simply focus on the much simpler, more concrete way of thinking, "Could random allocation alone produce this?" Although the tail areas are (one-sided) p-values we do not need the distractions of the p-value language and its related abstractions. We can keep the initial emphasis entirely on the concrete, "How does that compare to what random re-allocation produces?" But when the time comes when we do need to introduce these abstractions and their related jargon, so that we can enlarge the sphere of applicability of testing ideas, then we have an extensive array of already-familiar examples on which to ground them.

Additionally, we do not need to know about checking distributional assumptions and we do not need to worry about whether sample sizes are big enough for the calculations to give valid results. In the randomisation-test context sample-size considerations only become relevant when we introduce notions of test sensitivity or power.

6. DISCUSSION

We started with a challenge to statistics educators to find ways to get further into the rapidly exploding world of data faster and with better comprehension. We believe we have made some progress on a small part of this. In this paper the focus has been on early experiences of "statistical inference". Our iNZight data-exploration software project (Section 1.1) pushes out the boundaries in other directions. The approach we are targeting for introducing the thinking behind the basics of inference is to convey and exploit bootstrapping and randomisation as simple unified ways of thinking that give rapid transfer across a wide array of applications – sidestepping the standard slow progression through the cases: one sample means, one sample proportions, two sample means, two-sample proportions, k-sample means, …. Instead we wish students to be able to see a whole lot of closely related things almost at once. We are also very interested in increasing the accessibility of data exploration and inferential ideas to wider audiences. A key facilitator, we believe, is targeted visualisation software.

To help us think our way through the issues we have introduced and modelled the ideas of principled argument involving conceptual analysis. We advocate that these forms of discourse become common in statistics education as a way of facilitating community debate at fundamental levels, particularly about radical change.

As we noted in Wild et al. (2011), many of the problems with students learning statistics stem from too many concepts having to be learned and operationalized almost simultaneously pointing to the need for complexity-reduction strategies. This paper sits squarely in that space. We were interested

in making the basics of confidence intervals and significance tests accessible to as wide an audience as possible and also to be able to open up a wide range of important applications very quickly. Guiding principles were discussed at length in Section 2. These considerations led to our seeking the most concrete incarnations of these inferential ideas we could find and stripping them down to their barest bones for first experiences; this in order to avoid the trap of biting off too much and then losing what matters most to cognitive overload.

Our emphasis was fostering learning visually by establishing and working with mental images, i.e. things that can be imagined. The Holy Grail is images that are simple and vivid enough that they will tend to pop up automatically from the subconscious later in life at the times when they are needed (abiding images). Primarily, we have worked with dynamic visual images. The main point of the animation visualisations is motivation and demystification; building an appreciation that "this does sensible sorts of things". Psychology is important. "I roughly understand how these things work" should, if nothing else, help facilitate sensible choices and interpretations in applications. All of the visualisations emphasise observing random behaviour –showing random processes unfolding over time and leaving behind them a history that can be used for statistical reasoning. They also do double-duty as analysis tools for use in early experiences, tools that permit probing into how something was constructed and not just providing the end product.

The basic principle for conceptual pathways was keeping it as simple as possible until the basic ideas are well established. This already gets us quite a long way and for some audiences what is discussed here may well be sufficient. But for anyone going further in statistics we have to think in terms of this being a base camp from which to launch assaults on higher levels of the mountain. "Now that we have seen all this at work using the simplest versions of ideas, we will now dig further in to understand things more deeply, improve our methodology" and so on. There are generalisations that are easy because they demand very few new ideas and generalisations that demand a substantial investment in new conceptual foundations.

Easy generalisations include generalisations to more complicated sampling or experimental designs. These introduce only a strengthening of the underlying principle of the random mechanism used in analysis paralleling the random mechanism used in data generation. So with a randomised block design the random relabeling has to be done within blocks, complex survey-sampling mechanisms in data collection have to be mimicked in bootstrap re-sampling, and so on. Information about any complex randomisations that have been used in a study design *has always* to be supplied by the data analyst to a computer program, no matter what the inferential paradigm, in to order get a valid analysis. But while there are important knowledge bases around these elements they are adding nothing to the basic ideas of bootstrap confidence intervals or randomisation tests beyond "mimic the way it was done".

Generalisation to parameters beyond those displayed in Table 1 are immediate once someone both understands the new parameters and has acquired confidence in the use of bootstrap resampling distributions or randomisation distributions as general principles. What is difficult is to convey much beyond Table 1 visually in a way that makes appropriate connections between all of the elements. We do not think this really matters as the main point of the visualisations is to offer a wide enough array of experiences to foster comfort with and confidence in these two general principles. What is inherently much more difficult is the jump to confidence intervals based upon inverting

randomisation tests. They involve at a very fundamental level many of the conceptual elements we were able to avoid in Sections 4.5 and 5.3.

In this paper we have been almost entirely silent on pedagogy; which is why important nodes in Fig 1(a) are greyed out. Conceptual pathways and visualisation software are not pedagogy but simply aids to pedagogy. Conceptual pathways are like blue-print drawings – enormous amounts of knowledge, talent and effort are required to use them and create a great building. Visualisation software does not teach. Learning aided by visualisation software is a teacher-mediated process. This paper comes from the front-end thinking and software development for a substantial research project. The larger project has been a collaboration involving four statisticians, two statistics-education researchers, 16 high school teachers, seven university teachers and three teacher professional development facilitators and has involved development of activities and lessons, trialling and improvement of these in action-research cycles, and qualitative and quantitative research on large numbers of students. Other aspects of this research and its implications will be reported elsewhere, cf. the relationship between Wild et al. (2011) which concerned lower-level conceptual pathways, Pfannkuch et al. (2010) which discussed pedagogical dimensions connected with the use of language; and Arnold et al. (2011) which discussed pedagogy more broadly – including discussion of hands-on activities with the primary purpose of making what the computer is showing make sense. Simply watching an animation unfold is seldom, if ever, going to make something make sense (Wild, 2007). At a bare minimum, students have to be challenged to explain what the key elements are of what they are looking at, to describe what is happening, to explain why it is like that, and to discuss the lessons they have extracted from what they have been seeing. The job is not done when the teacher has explained it all to the student. The job is only done when the student can explain it all to someone else.

REFERENCES

Aldhous, P. (2011). Culturelab: The art of data, *New Scientist*, 5 February 2011, p44.

Arnold, P., Pfannkuch, M., Wild, C., Regan, M., & Budgett, S. (2011). Enhancing students' inferential reasoning: From hands on to "movies." *Journal of Statistics Education*, 19(2).

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

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(01):87–114.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. London: Chapman and Hall

Ernst, M.D. (2004).Permutation methods: A basis for exact inference. *Statistical Science*. 19(4), 676–685.

Forster, M., & Wild, C.J. (2009). Writing about findings: Integrating teaching and assessment. In P. Bidgood, N. Hunt, F. Jolliffe (Eds.), *Variety in Statistics Assessment.* New York: Wiley-Blackwell.

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

Hesterberg, T., Moore, D.S., Monaghan, S., Clipson, A., Epstein, R. and Craig, B.A. (2007), Bootstrap Methods and Permutation Tests, Chapter 16 for *Introduction to the Practice of Statistics*, 6th edition, by David S. Moore, George P. McCabe and Bruce A. Craig, W. H. Freeman, N.Y. [online: http://bcs.whfreeman.com/ips6e/content/cat_040/pdf/ips6e_chapter16.pdf]

Mayer, R.E. & Moreno, R. (2002). Animation as an aid to multimedia learning. *Educational Psychology Review*, 14(1), 87-99.

Clark, J. & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, 3(3), 149-210.

Petocz, A. and Newbery, G. (2010). On conceptual analysis as the primary qualitative approach to statistics education research in psychology. *Statistics Education Research Journal*, **9**(2), http://www.stat.auckland.ac.nz/~iase/serj/SERJ9(2)_Petocz_Newbery.pdf

Pfannkuch, M., Regan, M., Wild, C.J., & Horton, N. (2010). Telling data stories: essential dialogues for comparative reasoning. *Journal of Statistics Education*, 18(1), http://www.amstat.org/publications/jse/v18n1/pfannkuch.pdf

Pfannkuch, M., Arnold, P. and Wild, C.J. (2013). What I see is not quite the way it really is: students' emergent reasoning about sampling variability. Unpublished manuscript.

Smith, T.M.F. (1999). Discussion. *International Statistical Review*, 67(3), 248–250.

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cognitive Science*, 12(2), 257-285.

Ware, C. (2008). *Visual thinking for design*. Burlington, MA: Morgan Kaufmann Publishers.

Wild, C.J. & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry (with discussion). *International Statistical Review*, 67(3), 223–265.

Wild, C.J. Pfannkuch, M., Regan, M., and Horton, N.J. (2010). Inferential reasoning: learning to ``make a call'' in theory. *In Proceedings of the Eighth International Conference on Teaching Statistics* (ed. C. Reading). The Hague, The Netherlands: International Statistical Institute. Online: http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_8B1_WILD.pdf

Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. (2011). Towards more accessible conceptions of statistical inference (with Discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2), 247–295.