Using action research to develop a course in statistical inference for workplace-based adults

Sharleen Forbes
Victoria University of Wellington

Journal of Statistics Education Volume 22, Number 3 (2014),
www.amstat.org/publications/jse/v22n3/forbes.pdf

Key Words: Statistics education; Inferential statistics concepts; Hands on data and real world case studies; Use of exact probabilities; Action research.

Abstract

Many adults who need an understanding of statistical concepts have limited mathematical skills. They need a teaching approach that includes as little mathematical context as possible. Iterative participatory qualitative research (action research) was used to develop a statistical literacy course for adult learners informed by teaching in traditional first year university courses, workplace based training, teacher workshops and Masters of Public Policy courses. The latter learners in particular regularly come across confidence intervals and statistical significance in their everyday reading. The goal is to give them a conceptual rather than theoretical understanding of inferential concepts by developing inferential statistics logic through the introduction of exact probabilities in simple non-parametric tests (two-tailed coin tossing) and then contingency tables and parametric situations. The final course developed for the New Zealand Certificate of Official Statistics uses "hands-on" examples to reinforce concepts before proceeding to computer simulations. It emphasizes evaluation of the strength of statistical significance and its relationship to the possible cost of making an incorrect decision. Case studies that have influenced government policy reinforce inferential concepts and demonstrate the importance of statistics in complex real problems.

1. Background

My interest in approaching the teaching of statistical inference through non-parametric tests arose over twenty years ago, in final lectures in first year service courses in statistics at Victoria University of Wellington where the Sign Test was introduced as an example of how all the

(parametric) tests the students had encountered through the year (such as the t-test, ANOVA, etc.) had the same form: hypotheses, test statistic, calculation of the probability of getting a value at least as large as given by the sample(s) **if** the conjectured (null) hypothesis was true and a decision about whether or not to reject the null hypothesis based on the sample data. Over several years, I received feedback from students commenting on how they had missed the consistent "pattern" of the decision-making across the different tests and how they would like to have seen the Sign Test earlier in the course. Shortly thereafter, I left the formal academic environment, teaching only in ad hoc courses for social science graduates and workplace-based adults (both groups often having limited numeracy skills) until returning to the academic environment in 2007 to mainly teach post-graduate public policy students.

In the interim there has been a plethora of research on the teaching of informal inference including a special volume of the Statistics Education Research Journal introduced by Pratt and Ainley (2008). However, only a little is focused on the use or development of statistical inference in the workplace. Indeed, Bakker Kent, Derry, Noss, and Hoyles (2008) state that their research "suggests that few non-graduate employees need to interpret results stemming from hypothesis testing or confidence interval estimation" (p. 131). However, this is not the case with the participants in the courses discussed below. These participants are school mathematics and statistics teachers, general staff from the national statistics office, government employees and students in public policy courses. The majority are graduates and they are all expected to be able to interpret statistical information as part of their everyday work. Policy advisors in particular are routinely called upon to read papers containing research results. Action research in the form of observation and experiment was undertaken with students in three courses (described below) in 2009 and 2010.

The results were then used to develop and modify learning steps in inferential logic in the review of a course within the New Zealand Certificate of Official Statistics. This Certificate is specifically designed for students currently in full-time public sector employment and comprises four taught courses totaling 24 credits and a workplace based project totaling 16 credits. The particular course re-developed as a result of this research was called *Understand and evaluate sample survey techniques, estimation and hypothesis testing* and was worth 4 credits. The new version was taught for the first time in 2012.

2. Reducing the mathematics content in statistics courses

Cobb (2007) states that "we put the normal distribution, as an approximate sampling distribution for the mean, at the centre of our curriculum, instead of putting the core logic of inference at the centre" (p. 4). As our access to technology increases, so does our reliance both on formulae, and the need to understand the mathematics behind these formulae decrease. If we wish to use the normal distribution, we can justify visually that it works generally using simulation on readily available applets (such as http://onlinestatbook.com/stat_sim/sampling_dist/index.html) or packages like CAST (Stirling 2010). Re-sampling methods such as randomisation provide a viable alternative for calculating *p* values as shown, for example, by Rossman (2008) who demonstrated the use of randomisation with contingency tables as an alternative to the use of the approximate (Chi-squared) distribution to calculate *p* values. The question that statistics teachers now need to ask is whether it is more challenging, magical or mystical for students to be shown

visually that either re-sampling or the normal distribution of the sample mean works in practice, or to accept as given that the daunting (albeit beautiful) mathematics of the Central Limit Theorem is correct. Bootstrapping and randomisation are just being introduced to the senior secondary school curriculum in New Zealand (Ministry of Education 2012a), and software is currently being developed to make this accessible to all. As Cobb (2007) said we are becoming free of "the tyranny of the computable" (p. 4).

Teachers are now free to focus on developing the underlying statistical concepts, of which possibly the most important is probability. However, in the meantime, statisticians have done such a good job of "selling" t-tests, 95% confidence intervals and the use of 5% or 1% levels of statistical significance to the sciences in general that these now dominate the literature many educated adult learners are required to interpret. Progress may be slow in changing this focus, so care needs to be taken that new approaches to the teaching of estimation and inferential statistics do not create confusion for learners. Regardless of the methods used, we can persuade statistics learners to look at the strength of the quantitative evidence with which they are presented. This means putting the purpose or question we are asking up front and evaluating its importance in terms of the cost, either monetary or on the quality of life, etc., of making a wrong decision. This is not a new idea. As long ago as 1986 Speed (1986) stated "In my view the value of statistics, by which I mean both data and the techniques we use to analyse data, stems from its use in helping us to give answers of a special type to more or less well defined questions….yet I believe that much of the teaching of statistics and not a little statistical practice goes on as if something quite different was the value of statistics. …..graduates who find themselves working in government or semi-government agencies, business or industry, in areas such as health, education, welfare, economics, science and technology, are usually called upon to answer questions, not to analyse or model data, although of course the latter will be part of their approach to providing the answers" (p. 18). The level of importance attached to the question to be answered should be directly related to the level of significance (the probability of concluding that a change or difference exists when it does not) that we think is appropriate to use. This, of course, is not independent of the sample size, and there is debate (e.g., Ellis 2010) that we should concentrate on the *effect size* rather than the level of significance (*p value).* Although, as Ellis states, the "*p* value is a confounded index that reflects both the size of the effect and the size of the sample" (p.49), it is still the most common inferential statistic that the learners in the courses described here will be required to interpret. Ellis also states that interpretation of the practical significance or meaning of research is a social responsibility of scientists but it is rarely explicit in published research. It is, therefore, often left to readers to gauge the quality of the available data, right through from understanding variable definitions and categories (metadata), to the adequacy of sampling frames, sample and questionnaire or experiment designs and to the interpretation of analytical results. Thus, these are also skills that students need to acquire. When mathematical statistics is the focus of a course, care is needed to ensure it does not assume an unwarranted importance. For the learners described above, it is merely the tool that is used to estimate population parameters, test hypotheses or examine relationships between variables. The final part of the statistical process is the presentation or communication of results and increasingly, even for official statistics, this is delivered through new forms of dynamic and interactive graphics and maps (Forbes, Ralphs, Goodyear, and Pihama 2011) which students also need to be able to interpret.

3. Purpose of the study: Course development

The overall objective in this iterative study was to develop a course for adult learners involved in high-level decision-making that covered the aspects described above and equipped them with tools for making decisions. The particular focus of this study was on developing their skills to use and interpret statistical inference. The course was based around three aspects:

1. Developing a conceptual understanding of inferential statistics from a probabilistic rather than a mathematical basis, in particular using "hands-on" examples that would be familiar to the learners
2. Investigating the application of these concepts in complex real world problems through the use of case studies in the classroom, and
3. Discussing the strength of statistical evidence needed for the real world decision making faced by the learners.

Previous research that relates to each of these areas is summarised briefly below.

3.1 Probability: The base of statistics inference

There is evidence that even young children can demonstrate probabilistic reasoning. In their study of four year-old children Denison, Konopczynski, Garcia, and Xu (2006) suggest that "even in the absence of formal instructions, children may already have certain intuitions that allow them to use probability in their reasoning" (p. 5). This study, and that of Pange and Nikiforidou (2007), show that even pre-school children have elementary ideas of randomness and chance. In school, students meet the concept of proportion (fraction) long before they meet any measure of central tendency (mean, median or mode). For example, in the current New Zealand school mathematics and statistics curriculum there are published national standards that state after two years at school (age 7 to 8 years) students will be able to "find fractions of sets, shapes and quantities" and "describe the likelihoods of outcomes for a simple situation involving chance, using everyday language" (Ministry of Education 2012b). New Zealand's adult learners of today may not have experienced a curriculum as advanced as the current one, but it is likely that they will have experienced both the concepts of fraction and of probability in primary schooling. In adult life, the proportion concept is reinforced by much of the media reporting of statistics, whether this be the proportion of different ethnic groups in a national population determined from Census data or opinion poll results, and the probability concept through media reporting or direct experience with many games of chance, such as lotteries, horse racing and sport betting. It therefore seems logical to use proportions and probabilities first when trying to develop inferential logic. Biehler (2011) stated that "all our knowledge is uncertain" and queried how we, as teachers, should go about "building quantitative knowledge about uncertainty." While he was raising this with respect to school children, the issue also arises when teaching adult learners. That is, what and how many stages of learning are required to build up ideas of uncertainty to the level of "objective expression that statistics provides" (Biehler 2011). Helen McGillvray (2011) stated that "Probability is at the heart of statistics teaching" and the establishment of a probabilistic base for inferential statistics has been advocated by a number of authors. For example, Lindley (2011) suggests we "teach one concept that will embrace the whole of worthwhile statistics....That concept is probability" (p. 283).

The adults taught in all the courses in this project were, in the main, highly skilled government employees but with very mixed numeracy skills. For these students, as with many other practitioners, "the traditional road to statistical knowledge is blocked, for most, by a formidable wall of mathematics" (Efron and Tibshirani 1993, preface). For this reason the "hands-on" tasks were deliberately chosen so that they are simple, visual, familiar to the students, common across their varied workplaces and do not require new mathematical skills. Many will be familiar to school classroom teachers. All the students were competent computer users and it could be assumed that they would all have desktop access to a personally dedicated computer in their work environment. In all the courses discussed in this study, the introduction of a new concept, whether this be sampling, inference or regression, began with a hands-on classroom activity. There was then rapid progression to the use of the internet. For example, the mean was displayed as the "balancing point" or centre of gravity of a set of data by placing small (Lego) blocks, and changes in the standard deviation (as a measure of the thickness of "spread") was demonstrated by moving blocks from being clustered about the mean to having a wider spread. Thereafter, software such as EXCEL was used to generate values of the mean and standard deviation, and packages such as CAST (Stirling 2010) used to investigate distributional properties. Pfannkuch, Forbes, Harraway, Budgett, and Wild (2013) have since further developed this teaching strategy stating that students "should have hands-on simulation activities before moving to computer environments" (p.3). In all the hands-on examples used in the courses described below, the students are playing with data in a way similar to the PPDAC (Problem, Plan, Data, Analysis, Conclusion) cycle described by Wild and Pfannkuch (1999).

3.2 Use of real data

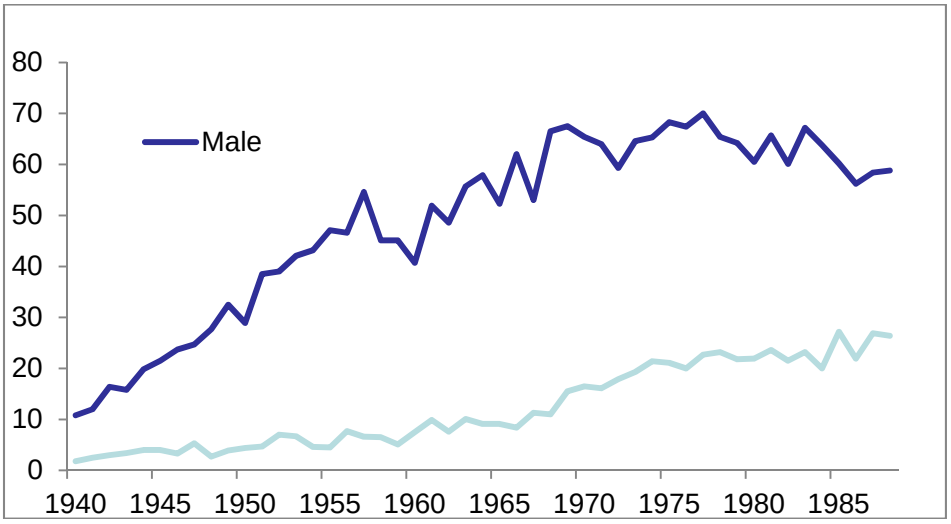
There is a large and still growing body of research supporting the use of real data in statistics education (from Dunkels 1994 through to more recent papers at the 2011 International Association of Statistics Education conference such as Crawford and Marriott 2011; Easterling 2011; Gal and Ograjensek 2011; and Payne 2011). Real data comes in many forms, the simplest being that generated in the classroom as, for example, described in the hands-on activities above. Real data is also generated in the workplace, and the students in the Certificate of Official Statistics will be required to use some of this data in their own projects. In many of my courses, some adult students steadfastly avoid learning material until its usefulness has been clearly demonstrated through such real world examples, so I make extensive use in my teaching of New Zealand case studies such as research reports and media articles that demonstrate the use and importance of statistical information in government decision-making. In the courses described below, real world data was presented both through on-line survey results and media or research reports containing inferential statistics.

3.3 Strength of statistical evidence in decision making

There has also been much discussion about practical versus statistical significance and desired sample sizes, etc. (e.g., Ellis 2010). In the real world, statistical questions are not all of equal importance. Some classroom examples may have little social or economic value and some of the case studies in the 2009 and 2010 Masters of Public Policy courses were chosen to demonstrate how statistical thinking contributes to decision-making about sometimes extremely complex and intractable problems. For example, one case study used to reinforce simple descriptive statistics

concepts (graphics, proportions and rates) and simple linear regression was the history of smoking policy (Appendix A). This shows how statistical information has resulted in a 180º turn around in New Zealand government policy from the support (almost promotion) of smoking through the government supply of cigarettes to the armed forces and prisoners from at least the First World War onwards (Diehl 1969 cited in Thompson and Wilson 1997) until the introduction of the New Zealand Smoke-free Environments Act in 1990 and subsequent amendments. This example demonstrates the current complex mix of anti-smoking incentive and disincentive policies, regulations and legislation. Presentation of specific statistical analyses that have influenced government, such as the post-war lung cancer mortality rates given in Figure 1, was used to discuss the different graphical displays used to "tell stories." One of the reasons for using this example was to demonstrate the high costs (in terms of individuals' health, medical costs and lost productivity) of wrong policy and of rectifying wrong policies (in terms of producing educational campaigns and enforcing new regulations and legislation).

Figure 1. Lung Cancer Mortality Rates
(*Department of Statistics and Department of Health 1992*)



4. Study method: An Action Research approach

In adult education, action learning or participatory qualitative research is based on the premise that "adults should have control over the content and form of their education" (Pant 2006, p. 95). In its strict sense, this is difficult to achieve. In only one of the classroom applications discussed in this paper were all the participants informed that it was a pilot/research project and invited to give feedback on both the content and the teaching methods used. However, as Cornwell and Jewkes (1995) state, "What is distinctive about participatory research is not the methods, but the methodological aspects of their application" (p. 1667). This is particularly the case when trying to teach and research simultaneously. Pant (2006) stated in reference to adult education that "maintaining a strict separation between the researcher and the subjects is also problematic" (p. 93) and, indeed, is almost impossible in any situation where there is also a teacher-student relationship in addition to the researcher-subject relationship. In the work reported here, the

researcher was both a participant (as the teacher) and a researcher/ learner. The research process was iterative over a period of several years and adaptive, with small changes being introduced over a number of courses. This is not unusual in participatory education research in that "participatory methodologies are often characterized as being reflexive, flexible and iterative, in contrast with the rigid linear designs of most conventional science" (Cornwell and Jewkes 1995, p. 1668). This project could be viewed as the type of participatory research described by the International Council of Adult Education as "an integrated activity that combines social investigation, educational work, and action" (cited in Pant 2006, p. 97). The teacher/ researcher was both gaining understanding and knowledge and taking action to make beneficial changes for present and future students. Cunliffe (2010) describes this process as "inter-subjective" with the researcher being involved in an "intricate flow of complexly entwined relational and responsive activities." She calls this type of research "abductive" with the researcher looking for "interpretive insights." Bakker et al. (2008) describe abduction as a method of reasoning whereby inferences are formed that may explain the observed data. While this research is not design-based, it is based on the conjecture that statistics education pedagogy can be improved for the students' benefit by informed consideration of what works in the classroom.

4.1 The students and the courses

Three different classroom situations are discussed as follows (in chronological order):

- an informal teachers workshop, *Understanding the logic of statistical inference,* for approximately 20 senior secondary school mathematics teachers attending a day organised by the Auckland Mathematics Association, in June 2009.
- a teaching experiment with a group of seven adults in the Christchurch branch of New Zealand's national statistics office, Statistics New Zealand in August 2009. The course called *Informal Inference: Making statistical decisions* was advertised as a "workplace pilot" and the promotion material contained that "the focus will be on the decision-making logic rather than learning formulae." The students were made aware that this was part of a research project and that although they were not going to be formally tested, their feedback was important. The students were not from the core statistical areas of the department but from service areas such as Information Technology, Data collection, and Processing, and each had access to a computer on their desk top during the course. Their numeracy level varied but none had studied statistics at a tertiary level. The course was five hours long and covered: Useful statistical measures (proportions, means and medians), Useful graphs, and Statistical judgments.
- modular (one or two day) executive *Masters in Public Policy* courses in 2009 and 2010 for the Australian and New Zealand School of Government (ANZSOG). The students in these courses are mainly middle tier state sector managers with varied statistical backgrounds. The teaching of New Zealand students in ANZSOG is not constrained in how it is delivered but is constrained by the need to cover the same material as that given to the Australian students. The major author for this *Evidence-Based Policy Making* course is George Agyrous of the University of New South Wales and much of the content is based on his recent book (Argyrous 2009). In both the 2009 and 2010 ANZSOG courses, the order of topics was descriptive statistics, sampling, estimation, inferential statistics then correlation and regression (covered in a two day intensive course).

The inferential teaching material in each course was adapted from information collected from previous course(s). Formal assessment of student learning was only undertaken in the Masters courses, but marks were not recorded for separate questions so the impact of the new probability based material in the inferential statistics section could not be measured separately. As stated earlier, the amount of mathematics in each course was limited to be as little as possible, but the first section of each of the latter two courses did include descriptive (summary) statistics, tables and graphs, and revision of proportions and percentages.

4.2 Data collection

Inferential logic was developed from a probability base by setting up simple hypotheses about proportions and testing the validity of these hypotheses by taking trials. It was first introduced in all the courses with:

1. A simple sign test looking at the proportion of heads (or tails) in a coin-tossing experiment using a two-tailed coin. The exact probability of getting all tails (if the coin was fair) was calculated after each toss then students were asked what they thought the outcome would be on the next toss. In a "Bayesian-like" approach, the students were using the probability of getting their particular result to make a decision about what was going on "back in the real world" of the coin (using the terminology of Arnold and Pfannkuch 2010). Using a two-tailed or unfair coin when teaching the binomial distribution is not a new idea (e.g., Maxwell 1994; Marchini 2004; Dunn 2005) and simulation Applets for this also exist (Schneiter 2008), but its use to look at the errors associated with inferential statistics may not be common. In these courses, the students were also asked at what probability level they would decide that the coin was unlikely to be fair.

One or more of the following exercises were then introduced:

2. A two-colour experiment using a box containing unknown proportions of blue and yellow "Post-It" papers constructed by using random numbers to stack the papers in a random order so that only the colour of the top paper was visible to students. In the *Teachers workshop,* the two-colour box contained 1/3 blue and 2/3 yellow "Post-It" papers. This experiment was also done in the 2010 *Masters of Public Policy* course and was adapted slightly so that the actual proportions (11/30 and 19/30) were unlikely to be guessed. In both courses, students were asked to guess the unknown proportions of the two colours in the unseen stack. Five samples each of size 10 were selected, then the exact probability of getting a value at least this far from the guessed (expected) value was calculated using the binomial distribution on the internet (http://stattrek.com/tables/binomial.aspx). The five samples were then combined to form a larger sample. If the original guess (hypothesis) was rejected, this process was repeated with a new guess.

The probability level at which students made decisions in the first two exercises was recorded in all the courses.

3. Small contingency tables, initially constructed using data generated in the classroom and discussed in terms of the apparent strength of the relationship between two categorical variables, were followed by a formal measure of the strength of this relationship using an example (Figure 2) developed by Agyrous (2009) that used (Goodman and Kruskal's) lambda (λ) together with the guide given in Table 1. The relationships in each of the tables in Figure 2 were discussed. For example, in the top table in Figure 2, there does appear to be a relationship between an employee's sex and salary, with women tending to receive lower salaries than men, but this is not a strong relationship when compared to that shown in the bottom table. The relative strength of the two relationships is shown by the values of λ (0.21 and 0.61 respectively). The discussion was then turned from measurement of strength of a relationship to the sampling situation and analysis of the statistical significance of a relationship using the learning steps demonstrated in the example in Figure 3 (reproduced from slides used in the classroom).

Table 1. Relative strength of association using λ:

Range (+/-)	Relative strength
0	No relationship
0 < 0.2	Very weak, negligible
0.2 < 0.4	Weak, low
0.4 < 0.7	Moderate
0.7 < 0.9	Strong, high, marked
0.9 < 1	Very high, very strong
1	Perfect relationship

Figure 2. Contingency table – income by sex example

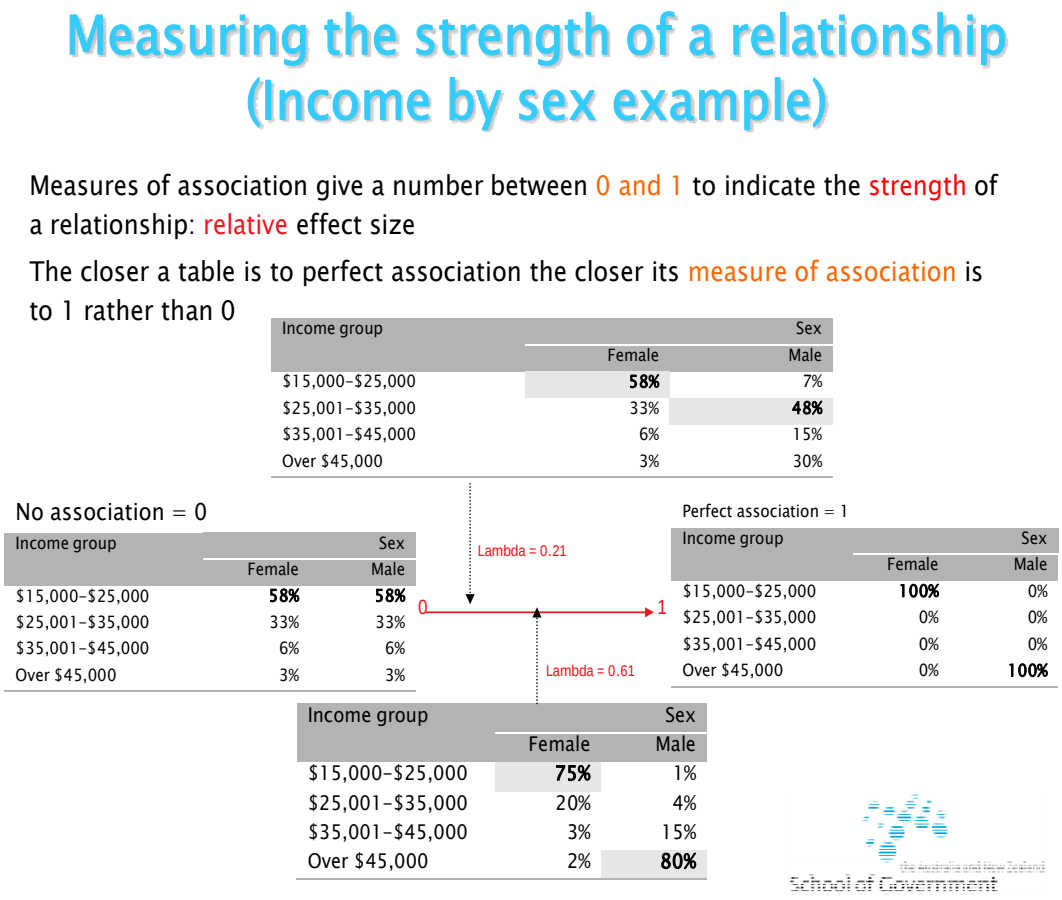


Figure 3. Contingency Table: TV preference by sex example

Simple bivariate analysis: Comparing frequencies

Two categorical variables – cross tabulations

Example 1:

I have a sample of 20 males and 30 females

• asked whether they preferred watching rugby or netball on TV.

Question: Is there a relationship between sex (gender) and preference?

Our (null) hypothesis is that there is no relationship.

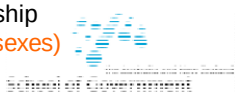
What is the call back in the population for each of the following samples?

	Rugby	Netball	Total
Male	20	0	20
Female	0	30	30
Total	20	30	50

Definite (perfect) relationship

	Rugby	Netball	Total
Male	10	10	20
Female	15	15	30
Total	25	25	50

Definitely no relationship
(equally split across sexes)



Categorical variables – cross tabulations

Example continued:

But what is the call for the following sample?

	Rugby	Netball	Total
Male	15	5	20
Female	10	20	30
Total	25	25	50

How far away from the expected table (if there is no relationship) is it?

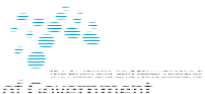
	Rugby	Netball	Total
Male	10	10	20
Female	15	15	30
Total	25	25	50

	Rugby	Netball	Total
Male	+5	-5	0
Female	-5	+5	0
Total	0	0	0

Take Sum (Differences^2) = $\frac{5^2}{10} + \frac{5^2}{10} + \frac{5^2}{15} + \frac{5^2}{15} = 8.333$

Expected

Too large???



The students were all competent computer users and were happy to be given the appropriate distribution (Chi-squared) and instructions on how to use the internet, as follows:

- Work out the *p* value for getting a value as large as our sampled value when there is really no relationship between the two variables by using the Chi-squared distribution at http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html
- Just input the (test) statistic [=8.333'] and the size of the table = (rows-1)x(columns-1) = [(2-1)x(2-1)] =1x1 =1, then press 'calculate.'

4. Playing with real data**.** In both years I introduced the concepts of samples and populations through a small hands-on task. Each learner was given a small bag of ten coloured lollies (Jelly-beans) and asked to work out the percentages of each colour. This was repeated in groups of four or five students (depending on class size). Group samples were then totalled to give the percentages for the whole class (population). This was partly introduced to revise fractions and percentages, but students quickly identified that each sample gave different proportions and that larger sample sizes were generally more likely to give better estimates of the population parameter. In both the *Workplace pilot* and the *Masters of Public Policy* courses, the distribution of a sample mean was introduced using small samples of real data collected in a class questionnaire used in the sample design session. A similar exercise was done to look at the characteristics of the sample proportion. Both produced histograms that looked roughly "normal." "Hands-on" data was also used to introduce the concept of association for numeric variables following a "living" example adapted by Mike Camden of Statistics New Zealand from Joiner (1975). The students were asked to stand and take their pulse, record it, then follow the teacher quickly up and down a flight of stairs and take their pulse again. The students then formed lines according to whether their initial pulse rate (at rest) was between 50-59, 60-69, etc. (a "Joiner" living histogram), then arranged themselves along each bar according to the value of their second pulse rate (after exercise) to form a living scattergram**.** A best-fitting living line was then "drawn" with a long piece of rope.

Computers were only available for students' personal use in the classroom in one of the courses, the *Workplace pilot.* In this course, the students accessed and played with real data in the form of random samples taken from the New Zealand CensusAtSchools (C@S) database (http://www.censusatschool.org.nz/; current url, http://new.censusatschool.org.nz/). This data was input by school children in response to questions about their favourite learning areas and sports, cell phone use, mode of travel to school, weight of school bag, etc. The students used the data to explore whether or not there was a (statistically significant) difference between the medians of two populations of school children.

The language of *making a call* from the sample back to the population developed by Arnold and Pfannkuch (2010) in their work for New Zealand secondary schools was used in these discussions. Prior to playing with the CensusAtSchool data, students were given simple rules for making decisions about medians using boxplots based on those developed by Wild, Pfannkuch, Regan, and Horton (2011) and also simple rules for making decisions about means based on the degree of overlap of confidence intervals (Figure 4). These rules could be seen as a first-step application of the forest plot pictures developed by Cummings (2006) in his meta-analysis work. In the courses described here, the students were told that, **if** the sample sizes are large (n>30) and roughly the same **and** the sample variances (spreads) are similar, then we can compare the 95%

confidence intervals for each mean using these "rules of thumb." It was suggested that the students should just use one of these sets of rules and could choose which level that they used.

Figure 4. "Rules of thumb" for making decisions about differences between population means at the 5% level of significance

95% Confidence Intervals	**Overlap**	**Making the call (5% level)**
(-------+-------) **(-------+-------)**	**None**	Means are **statistically different**
(-------+-------) **(-------+-------)**	**Only a little**	Means are **statistically different**
(-------+-------) **(-------+-------)** **(--------+--------)** **(--------+--------)**	**Less than half the width**	**What is the call here?** **Rule:** • If overlap less than 1/3 distance between the means then **different** • If overlap greater than 1/3 distance between the means then **not different**
(--------+--------) **(--------+--------)**	**More than half the width**	Means are **not statistically different**
(--------+--------) **(--------+--------)**	**Completely overlap**	Means are **not statistically different**

5. Use of case studies**.** Discussion of significance tests used in case studies was then used to reinforce the students' interpretation of *p* values. These were extensively used in the *Masters in Public Policy* courses where more teaching time was available. Examples of the case studies used are given in section 6 below.

Different forms of interaction between the teacher/ researcher and the students took place in each course, but in each, the teacher/researcher recorded brief notes both during and after classes on students' reactions and interactions, including areas that they found confusing. Students' evaluations (both formal and informal) were taken into consideration in making adjustments to teaching materials between courses.

While the purpose of this research was interpretive insight, in the sense described by Cunliffe above, the analysis and reporting of the data obtained in the research clearly sits more within the sort of "objectivist" paradigm defined by Morgan and Smirch (1980, cited in Cunliffe 2010). That is, the analysis focused on the understanding and behaviour of the students as objects of the research and the results were then used to modify course content and teaching practices in future courses. Table 2 lists the activities undertaken in each of the courses.

Table 2. Teaching activities used in each course

Exercise Course	Two-tailed coin	Two-colour experiment	Contingency tables	C@S data	Living Scattergram	Policy and Media Case Studies
Teachers workshop	X	X				
Workplace pilot	X		X	X		
Masters in Public Policy	X	X	X		X	X

5. Results

5.1 The double-sided (two-tailed) coin problem

In the two-tailed coin tossing exercise, the students in all the courses knew that the chance of getting a head or a tail on each toss of a fair coin was ½. It appeared to be a revelation to some of these students, in the *Workplace pilot* in particular, that their underlying assumption that the coin was fair could be wrong. When class discussion was focused on at the probability level at which they would question their original assumption, the majority view was that it should be less than the p=0.05 that is in common use (Table 3). Informal feedback from the course participants in the latter two courses indicated that this was the first time many of them had realised that statistical significance was a "probability" call.

Table 3. Points at which decisions about fairness were made.

Two-tailed coin -	**Number of tosses before coin questioned**	**Probability of occurrence if coin fair**	**Consensus view of level that should be used to reject fairness of coin**
Teachers workshop	7	p=0.008	$p < 0.05$ Note: this would already be a familiar level to the teachers
Workplace pilot	10	p=0.00097 Note: Two students remained adamant that the next toss would be a head *'on the law on averages'*	Very much less than p=0.05
Masters in Public Policy	8	p = 0.0039	p value = 0.004 (Note: is the level at which questions were raised)

5.2 Two-colour experiment

The *Workshop teachers* were familiar with both the binomial distribution and the idea of taking more data if samples are small. However, the *Masters of Public Policy* students appeared to be equally comfortable with the use of the internet to calculate the exact probabilities. Similar results were obtained in both courses, with the initial guess being p(blue=0.5). Table 4 gives the actual results from the *Teachers workshop* (using a one-tailed test for simplicity).

Table 4. Number of blue papers in each sample

Sample	**1**	**2**	**3**	**4**	**5**
No. of blue	3	4	2	2	4
***p* value**	0.17	0.38	0.05	0.05	0.38

The combined sample of 15 blue from 50 trials had a *p* value of 0.003. One of the teachers then said "this is only 3 chances in 1000 if there are equal numbers in the box," leading the others to quickly reject that p=0.5. A second guess of p(blue) = 0.3 was suggested and another 5 samples of size 10 were selected. Again, there was no clear decision on the basis of the five small samples, so these were combined to give 17 blue from 50 trials with a *p* value = 0.78. The teachers then decided that this was clear evidence that the correct proportion was 0.3. When I informed them that the actual proportion was not 0.3 but it was very close, one teacher immediately guessed the correct value of p=1/3. A similar situation arose with the *Masters of Public Policy* students except that they could not guess the correct proportion of blue papers.

The teachers in particular knew that each sample would give a different value but, in both courses, students were unsure whether or not to reject that p(blue) = 0.5 when faced with the somewhat conflicting *p* values shown in Table 4, and appeared relieved when I suggested combining the results to get a larger sample. There was also an immediate reaction of surprise and concern that the data "reinforced" an incorrect proportion, providing a natural progression to discussion about the errors (type 1 and type 2) associated with hypotheses testing. It seemed that this hands-on activity led to a more natural and animated discussion of these errors and reinforcement that not rejecting the null hypothesis does not imply its acceptance, compared to what is usually the case in more formal "chalk and talk" teaching. It also provided an opportunity for discussion about the lack of "power" of small samples. Two insights were gained by the teacher/researcher in these courses. One was that although these students could make inferential decisions when faced with just one *p* value, they were hesitant when faced with multiple *p* values, a situation that often occurs when researching the scientific literature on any particular topic. The other was that the two-colour problem could be extended to demonstrate the errors associated with inferential statistics and reinforce that not rejecting the original hypothesis is not equivalent to an acceptance that it is true. For some of the students in these courses, however, this required an undoing of previous teaching of *acceptance* of the null hypothesis.

5.3 Contingency tables

In both the *Workplace pilot* and the *Masters of Public Policy* course, students actively participated in constructing small contingency tables using data generated from their classmates

and had no difficulty in seeing the differences between those demonstrating weak and strong relationships. One possibility for the future is to reinforce the differences between the tables with pictures (such as the mosaic plots used by Unwin 2003). Introduction of a measure of association was only done in the *Masters of Public Policy* course. The students understood the different levels of strength given in Table 1, although from a teaching perspective, it was extremely difficult to convey to this group why we often have different values of λ when using variable A(say) to predict variable B than when using B to predict A. In the Certificate of Official Statistics, λ will be replaced by Cramer's C that has the advantages of being a single value, derived directly from the sum obtained above in Figure 3 (the Chi-squared statistic), and can be calculated (for small tables) by simply entering the cell values into the internet (e.g,. http://www.vassarstats.net/newcs.html). When used in the sampling situation, these measures belong to the *r* family of effect sizes described by Ellis (2010).

All the students in both courses indicated that they understood they were trying to measure how far away the values in the table were from those they would, on average, expect to get if the hypothesis of no relationship between the variables was true (my words). While they were all willing to accept the mathematics in the calculation of the test statistic in Figure 3, one of the seven in the *Workplace pilot* admitted that they "couldn't really see how we got to it," until the teacher slowly worked through it again with them.

The *Workshop pilot* students were able to use their desk-top computers to obtain the result that the probability of a sample giving a table at least this far away from our expected table when there is no relationship, was equal to 0.004 *(p* value), and decided that this was sufficient evidence to reject the hypothesis of no relationship. One student stated that they "were more interested in seeing if the answer (the *p* value) looked right than understanding all the mathematics" – an attitude that many applied statisticians would welcome. When questioned, the students indicated that they would have found it easier to just input the table into the computer (without calculating the test statistic) and obtain the *p* value. The teaching insight gained was that the mathematical construction of the Chi-squared statistic confused some students and interfered with the inferential decision-making process. There are some websites that enable this, such as the one used to derive Cramer's C above, and this was selected for use in the final Certificate of Official Statistics course. Within the classroom, there was no discussion about the Chi-squared distribution being only an approximation, and little about those situations (e.g., sparse tables) when it is not a good approximation. Students were instead given a simple set of rules when they shouldn't use it.

Consideration was given to using a randomisation procedure to estimate the probability of getting a table as far away from the expected table under the hypothesis of no association. That is, essentially a simulation procedure similar to the permutation test described by Pagano and Halvosen (1981), "Enumerate all possible tables consistent with the given marginals, and calculate the probability of each. The significance value of the observed table is then the sum of the probabilities of those tables that are as likely or less likely than the observed table" (p. 931). This would help reinforce the probabilistic basis behind the inferential decision, and remove the confusion introduced by the mathematical barrier of calculating the test statistics but is too time consuming for short courses.

5.4 Using real CensusAtSchools data

The *Workplace pilot* students used their desk-top computers in the classroom to take their choice of random samples from the New Zealand CensusAtSchools, following a demonstration using the heights of 11 year old male and female students to explore whether or not there was a (statistically significant) difference between the medians or means of the two populations. These students seemed very confused about the rules of thumb given, one commenting that they "hid the probability." The teacher used the internet to calculate the exact probability of getting a value as large as the sample difference if there really was no difference "back in the population," and the students were then told they could choose their own method of determining statistical significance. Most of their discussion centered around whether or not there was sufficient overlap of the sample boxplots to make a call back in the population. These students were all very enthusiastic about the use of the CensusAtSchool data, commenting that they "didn't want to stop playing with it" because they were so interested in the results. Almost all were parents, so this is perhaps not a surprising result.

In estimation, the level of confidence (significance) is set at 5%, and it appears that, from a learner's perspective, this contradicts the probabilistic logic behind inferential decision making, so neither set of rules of thumb were not included in the final Certificate of Official Statistics.

5.5 Living scattergram

This exercise was only used with the *Masters of Public Policy* students, originally as an activity for the after-lunch classroom lull. However, the students' enjoyment of the exercise, even from less physically active students, was carried back to the classroom where, after the equation for a line was presented and the interpretation of R^2 discussed, there was rapid progress from simple regression to interpretation of results in outputs from multiple regression (as presented in the case studies). No mathematical formulae were given for correlation coefficients, etc., and discussion focused on interpreting these. Students worked together in groups and on at least two occasions, were overheard using the "pulse" example as the basis of their discussions. One student commented that they would like to have the living regression exercise written down on paper as it helped them remember what the components of the regression line were.

5.6 Parametric tests

Although they could follow the examples given, a few of the students in both the *Workplace pilot* and the *Masters of Public Policy* courses remained somewhat puzzled during the discussions on the distributions of the sampling mean and sampling proportion. Students in all the courses were competent computer users, so this learning will be reinforced in the Certificate of Official Statistics by rapid transition to further simulations using the CAST software (Stirling 2010).

It was almost the same set of students who had difficulty with the standard deviation when summary statistics were taught who had difficulty with the concept of the standard deviation of the sample mean. However, most had no difficulty in interpreting the *p* values. In my view, they would have preferred to have just been told the distribution (or url) to use to calculate the *p*

values, as in the previous situations. While acknowledging these are slightly different situations, these students seemed to have more difficulty with the concept of "spread" than that of "correlation," and my conjecture is that this may in part be a language issue in that *co-relation* is a better descriptor (from a non-mathematical point of view) of the concept of correlation than standard deviation is of variation.

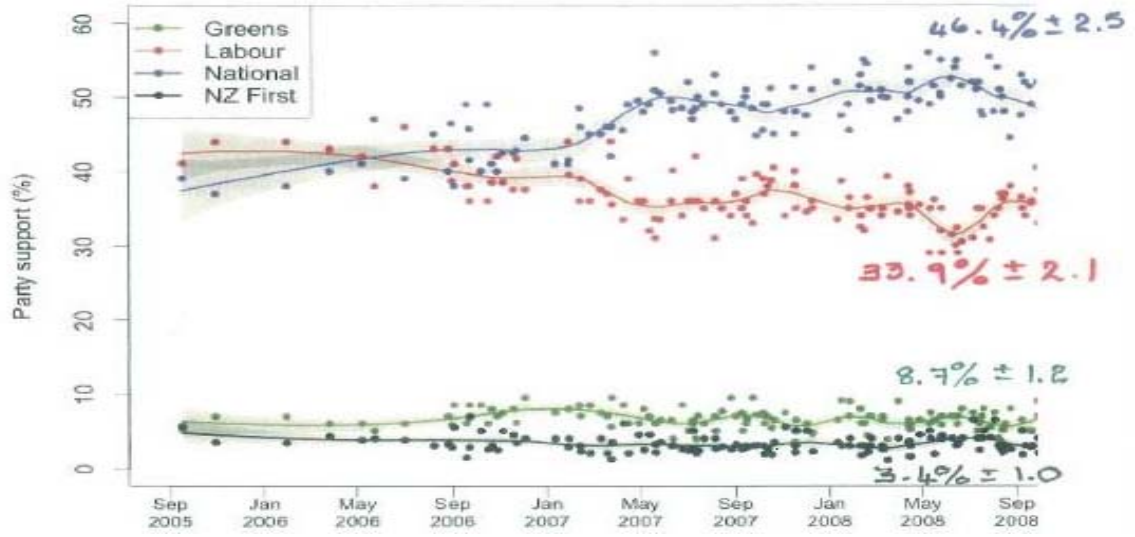
5.7 Case Studies

The use of case studies to reinforce the students' interpretation of *p* values was highly successful. In the *Masters of Public Policy* course in particular, students enthusiastically participated in small groups, interpreting and discussing what the results meant. The use of research reports, policy and media articles is extended further in the Certificate of Official Statistics.

6. Examples of case studies used to reinforce learning

In addition to the history of smoking policy example given above, the case studies used in 2010 included: statistical media releases, (such as the Household Labour Force Survey, Statistics New Zealand 2010); excerpts from Government surveys (such as the 2001 and 2006 Māori Language Surveys, Te Puni Kokiri 2002 and 2007); media articles (e.g., "Gore the least gay town in New Zealand," Gault and Chapple 2007); and research reports (such as Russell and Haney 1997). A case study of pre-election opinion poll data from Wikipedia (2008), as displayed in Figure 5, was used to discuss both time series and confidence intervals. The teacher's hand-written confidence intervals on this diagram are for the latest time point given.

Figure 5. Wikipedia display: Opinion polls prior to the 2008 New Zealand election.



Another case study related the outcome of research on the health and economic benefits of housing insulation for lower socio-economic families in New Zealand (Howden-Chapman et al. 2007) to the recent government policy of providing economic assistance towards housing insulation (retrofits). Originally, subsidies were offered only to low-income families, but in 2009,

this was extended to all families with houses built before 2000. This "Warm Up New Zealand; Heat Smart" scheme offers "up to $1,300 (or 33%) towards the cost of ceiling and under floor insulation if you are on a general income or up to 60% if you have a Community Services Card" http://www.eeca.govt.nz/node/3107. The Energy Efficiency and Conservation Authority (EECA 2011) reported that 163,700 homes were insulated under this scheme between 2009 and 2012, and that a study (Howden-Chapman cited in EECA 2011) estimated that the current rate of retrofitting translated to over 40 avoided deaths annually in the elderly population. They stated that "P. Howden Chapman's research…has significance" and that "health research as a whole has helped to justify the continuation and expansion of the retrofit programmes." The design aspects of this randomised control trial were discussed with the class, and Table 5 below was used to look at the strength of significance and also to clarify the distinction between the differences between the means and the mean difference. This case study was used to demonstrate the high cost of Government policy and emphasise the need to make the right decisions. A separate table from the same study was used to look at consistency of direction in social and health outcomes.

Table 5. SF-36 scores as reported in Howden-Chapman et al. (2007)

Table 7 | Self reported SF-36 results in a trial of insulating houses. Results are mean score in adults who had data for both years, unless stated otherwise

	Before intervention		After intervention		Difference (95% CI)	
Scale	Interven-tion group	Control group	Interven-tion group	Control group	Unadjusted	Adjusted
Social functioning	69.2	69.3	78.4	72.3	6.1 (3.9 to 8.4) P<0.0001	6.2 (3.8 to 8.6) P<0.0001*
Role emotional	63.1	62.4	77.5	66.7	10.8(7.2to14.5) P<0.0001	10.9 (7.1 to 14.6) P<0.0001*
Role physical	52.5	52.2	70.0	58.8	11.2 (7.4 to 15) P<0.0001	11.8 (8.0 to 15.5) P<0.0001*

*Adjusted for score at baseline, age group, sex, ethnic origin, household, and region.

Not only did the students enjoy the case studies, but, in my opinion the real world problems presented in these studies were a powerful motivating device, demonstrating to students the variety of ways that statistics are used as an input to policy advice. The most general change I observed in the students was an increased confidence in interpreting the statistical information presented in the case studies. As one student said with regard to the Howden-Chapman report, "even though I still don't know what all of the (statistical) terms mean I can see what the *p* values are saying." However, some students also queried why 95% confidence intervals had been calculated in the Howden-Chapman report when the *p* values were very much smaller than 5%, suggesting that we teachers still have some way to go to integrate estimation and inferential concepts.

7. The new course

The revised version of the Certificate of Official Statistics course, developed from the work described in this paper, was taught over two days in 2012, with both hands-on experiments and real case studies being used. In general, students worked together in small group workshops (a format used in all the Certificate courses). Examples of hands-on experiments, case studies and computer simulations used in this course are given in the course outline below:

Day 1 Morning – Conceptual frameworks and administrative data sets

- **Real case study** - Achievement at Maori-Medium schools (Wang and Harkess 2007) used to discuss timing and counting issues associated with administrative data

Day 1 Afternoon – survey samples

- **Hands-on experiment** - lolly sampling experiment to investigate random error and the effect of sample size on random error
- **Real case studies**
  - Statistics New Zealand's Retail Trade surveys (Statistics New Zealand 2012b) used to discuss different types of sample designs
  - "Effect of insulating existing houses on health inequality," Howden-Chapman et al. (2007) to discuss designed experiments

Day 2 Morning – Estimation

- **Computer simulation** of distribution of sampling means using CAST (Stirling 2010)
- **Real case studies** - Tongan Household Income and Expenditure Survey– to calculate confidence intervals
  - – "Opinion polling for the New Zealand general election, 2011" (Wikipedia 2011) - to use confidence intervals to look at differences between population means for both the major and the small parties

Day 2 Afternoon - Statistical inference and hypothesis testing

- **Hands-on experiment** Coin tossing example to develop inferential concepts and discuss Type 1 and 2 errors
- **Internet use**
  - *http://stattrek.com/tables/normal.aspx* to calculate *p* value for difference in means
- **Internet use**
  - *http://faculty.vassar.edu/lowry/newcs.html* to calculate *p* values for contingency tables
- **Hands-on experiment** - *collecting pulse before and after exercise followed by*
- **Computer use**
  - *http://www.danielsoper.com/statcalc3/calc.aspx?id=44* to calculate *p* value for correlation coefficient.
- **Real case studies** Repeat of – "Opinion polling for the New Zealand general election, 2011" (Wikipedia 2011) to calculate *p* values using the internet.
  - History of smoking example to look at the impact of statistics on policy development as discussed previously
  - "Effect of insulating existing houses on health inequality," Howden-Chapman et al. (2007) to discuss and interpret *p* values reported in research papers.

In addition to the classroom instruction, a full written workbook was supplied that contained urls to internet sites for the use of simulations and to derive *p* values, etc., that could readily be developed into an e-book. The assessment questions for this module were also based on two case studies, a survey of problem gambling in New Zealand (Arnold and Mason 2007) and the June 2012 quarterly release of the Household Labour Force Survey (Statistics New Zealand 2012a).

Two small cohorts of students took this course in 2012, fourteen in Wellington and eight in Christchurch. The courses were co-taught with an academic colleague from Canterbury University. Unfortunately, only half (50%) of the students filled in student evaluation forms, but the majority of these viewed the course positively. Ten of the eleven who responded rated the course overall as good or very good. Nine of the eleven agreed or strongly agreed that that the course structure was clear and logical and that the knowledge and skills gained in this course will be valuable for their work, eight that the content clearly related to the course objectives, and that the presenters encouraged questions and interaction and seven that they communicated ideas and information clearly. All the other responses were neutral apart from one respondent who clearly expected and wanted more formal statistical analysis of the type they had experienced in previous (first year university) courses.

8. Action Research insights

The particular insights gained in this action research were that:

- Students enjoyed, participated in and seemed to retain the knowledge from hands-on activities using their own data. This is one particular form of real data.
- It seemed to be easier for these non-mathematical students to accept, understand or understand and interpret the concept of association or correlation rather than that of spread as defined by the standard deviation.
- In the coin-tossing experiment, the adult students generally did not decide that their original hypothesis of a fair coin was incorrect until the probability of getting their expected result was very much smaller than the $p < 0.05$ in common usage. However, across all these types of classroom situations, there seemed to be general agreement that a few in a thousand was a small enough probability for rejection of the null hypothesis.
- In all the above different classroom situations, after calculating one or two exact probabilities themselves, students were happy to use the internet to determine these and readily accept advice on which test to use when. However, their feedback also reinforced that they like to see what was happening in the data. This gives some support to the use of randomisation in the classroom to reinforce the probabilistic base of inferential logic.
- Students found it much easier to make inferential decisions when faced with one rather than multiple *p* values (as often occurs in practice when reading research or accumulating evidence). Teachers could help reinforce that, as Ellis (2010) states, "The best test of whether a result is real is whether it can be replicated at different times and in different settings" (p. 49).
- Use of real world case studies about significant policy problems provides a natural segue to discussion about what level of significance should be used for what size "real world" problem and how important (in terms of cost) it is not to make a wrong decision.

9. Concluding Comments

In all types of classrooms (schools, university or workplace), it is difficult to set up even quasi-designed experiments that are perceived to be fair by all the participants. Indeed, it is almost impossible in courses without formal assessment, such as in this study, to provide quantitative evidence about whether or not students conceptual understanding of inferential statistics was greater if this was taught from a probabilistic basis than if it was taught from the traditional parametric approach. However, what is certain, is that the development of conceptual understanding takes time, and that the time constraints of many service courses don't allow for a phased approach to building understanding.

The type of adaptive participatory research used in this study, where the teacher learns from students' interactions, behavior or feedback, either over a number of courses or over a period of time, can provide valuable insights. Indeed, this integrated activity combining social investigation with educational work and action, is just a formalisation of the activities one would expect the best teachers to be doing in their ongoing work.

Using binomial tests and then contingency tables was an effective way of demonstrating that statistical decision-making has probability as its basis. Randomisation procedures could also be used to reinforce this. While in this study the idea of strength of significance was discussed in terms of its practical significance, or the cost of the real world problem, perhaps the next step is to introduce a hierarchy of "strength of evidence" such as that proposed by Wild and Seber (2000). That is, a p value < 0.1 = weak evidence; p value < 0.05 = some evidence; p value; < 0.01 = strong evidence and p value < 0.001 = very strong evidence. This would provide some consistency with the "strength of relationship" hierarchies that the students readily accept and use with respect to measures of association. It is acknowledged that the p value does confound both the size of an effect and the size of the sample, but it is still the most common measure presented in the research literature and is likely to be for some time so does need to be understood by students. We would probably all accept that the strength of statistical evidence needed in real world decision making should vary according to the cost or importance of the issue as was debated by these students. However, if we are asking students to look at, and relate the strength of evidence represented by the p value back to the importance of their real world problem, then why are we also expecting them to calculate intervals with set levels of confidence? Tables containing p values together with 95% confidence intervals are the norm in the scientific literature. There may be an argument for having a standard for comparative purposes, but should we not be giving decision-makers estimates that are of sufficient width to reflect the complex real world in which they are making multi-million dollar decisions? A question for statisticians is whether or not we have oversold to the rest of the scientific world both the 5% level of significance and the 95% confidence level especially now that we are free from computability constraints and, if so, how can we rectify this?

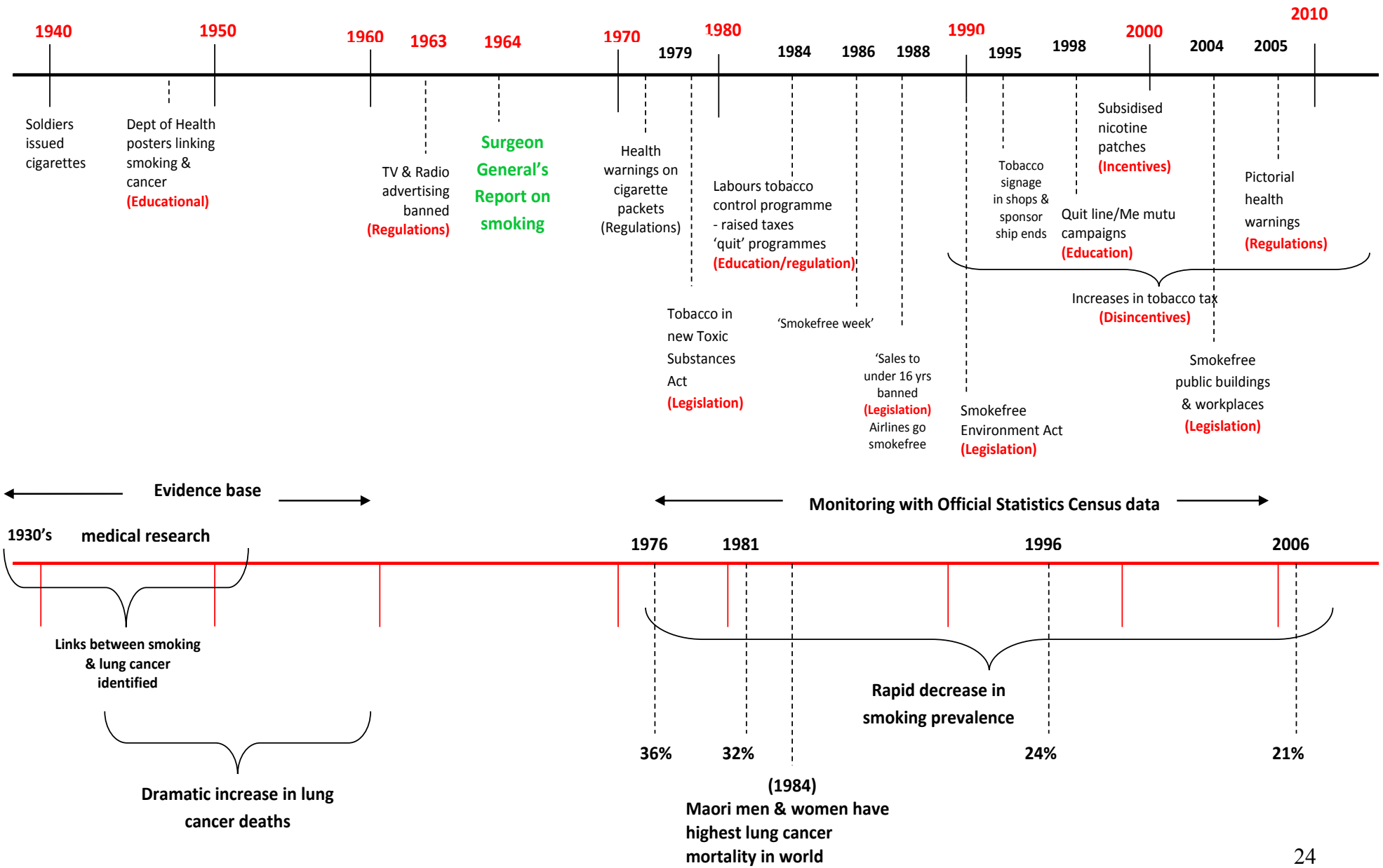
This study also provided further evidence of the value of using hands-on activities, real world data and real world problems in the classroom as shown in the enthusiasm for hands-on activities and use of real (CensusAtSchools) data and the interest shown in the case studies. It has become the practice in the Certificate of Official Statistics for any new concept to be introduced with a

hands-on exercise followed, where appropriate, by computer simulations and for real world case studies to be used as often as possible.

Appendix A

The Power of Statistics

History of Smoking Policy in New Zealand (major source: The Smokefree Coalition)



References

Argyrous, G. (ed.) (2009), *Evidence for policy and decision-making*, Sydney: UNSW Press.

Arnold P. and Pfannkuch, M. (2010), "Enhancing students' inferential reasoning: from hands on to 'movie snapshots'", *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS 8),* [online], Available at http://www.stat.auckland.ac.nz/~iase/publications.php?show=icots8

Arnold, P., Pfannkuch, M., Wild, C.J., Regan, M. and Budgett, S. (2011), "Enhancing students' inferential reasoning: From hands-on to 'movies'", *Journal of Statistics Education*, *19*(2), 32 pages, http://www.amstat.org/publications/jse/v19n2/pfannkuch.pdf

Arnold, R. and Mason, K. (2007), "Problem gambling risk factors and associated behaviours and health status: results from the 2002/03 New Zealand Health Survey", *New Zealand Medical Journal* [online], 120, 1257. Available at http://www.nzma.org.nz/journal/120-1257/2604/.

Bakker, A., Kent, P., Derry, J., Noss, R., and Hoyles, C. (2008), "Statistical Inference at Work: Statistical Case Control as an Example", *Statistics Education Research Journal, 7*, 2, 130-145.

Biehler, R. (2011), Discussant's presentation given to The Seventh International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-7), Utrecht University, The Netherlands, July, 2011.

Cobb, G. (2007), "The Introductory Statistics course: A Ptolemaic Curriculum?", *Technology Innovations in Statistics Education* [online],. 1, 1, Art. 1. Available at http://repositiories.cdlib.org/uclastats/cts/tise/vol1/iss1/art1.

Cornwall, A. and Jewkes, R. (1995), "What is participatory research?" *Social Science and Medicine*. 41 12, 1667-1676, Great Britain: Elsevier Science Ltd,

Crawford, E. and Marriott J. (2011), "Teaching Statistics through problem solving: using real time data retrieval," in *Proceedings of the 2011 IASE Satellite Conference: Statistics Education and Outreach.* ed. P. Bidgood [online]. Available at http://www.conkerstatistics.co.uk/iase/proceedings.php

Cummings, G. (2006), "Meta-analysis: Pictures that explain how experimental findings can be integrated", *Proceedings of the 7th International Conference on Teaching Statistics. (ICOTS7) Salvador, Brazil* [online]. Available at http:www.stat.auckland.ac.nz/~iase/publications/17/C105.pdf .

Cunliffe, A.L. (2010), "Crafting Qualitative research: Morgan and Smircich 30 years on", *Organizational Research Methods* [online]. Available at DOI:10.1177/1094428110373658. SAGE Journals. http://orm.sagepub.com/content/early/2010/07/23/1094428110373658.

Denison, S., Konopczynski, K., Garcia, V., and Xu, F. (2006), "Probabilistic reasoning in preschoolers: random sampling and base rate," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, eds. R. Sun and N. Miyake, Mahwah, NJ: Erlbaum, pp.1216-1221.

Department of Statistics and Department of Health (1992), *Tobacco Statistics 1991. Trends in Tobacco Consumption and Smoking Prevalence in New Zealand*, Authors, Wellington.

Dunkels, A. (1994), "Interweaving numbers, shapes, statistics and the real world in primary school and primary teacher education," in *Selected lectures from the Seventh International Congress on Mathematics Education.* Robitaile, D. F., Wheeler, D. H. and Keirnan, C. Quebec: Les Presses de l'Universite Laval.

Dunn, P.K. (2005), "We can still learn about probability by rolling dice and tossing coins". *Teaching Statistics*. 20, 2, 37-41.

Easterling, R. (2011), "Passion- Driven Statistics," in *Proceedings of the 2011 IASE Satellite Conference: Statistics Education and Outreach.* ed. P. Bidgood [online]. Available at http://www.conkerstatistics.co.uk/iase/proceedings.php

EECA (Energy Efficiency and Conservation Authority) (2011), Personal email to author in response to questions about the scheme and the influence of Howden-Chapman's research on policy (dated 27 October 2011).

Efron, B. and Tibshirani, R. (1993), *An introduction to the Bootstrap*. Chapman & Hall.

Ellis, P.D. (2010), *The essential Guide to Effect Sizes. Statistical Power, Meta-Analysis, and the Interpretation of Research Results.* Cambridge University Press.

Forbes, S., Ralphs, M., Goodyear, R., and Pihama, N. (2011), "New ways of visualising official statistics". Working paper series. No 20, Cairo, Egypt: Information and Decision Support Centre.

Gal, I. and Ograjenšek, I. (2011), "Using customer satisfaction surveys as a teaching resource. In statistics education: methods and benefits," in *Proceedings of the 2011 IASE Satellite Conference: Statistics Education and Outreach.* ed. P. Bidgood [online]. Available at http://www.conkerstatistics.co.uk/iase/proceedings.php

Gault, L., and Chapple, I., (2007), "Gore the least gay town in New Zealand", *Sunday Star Times*, Sunday, 29 July 2007. Auckland, New Zealand

Harraway, J. (2011), "Use of case studies and new software to motivate statistics teaching and learning at school and undergraduate level," in *Proceedings of the 2011 IASE Satellite Conference: Statistics Education and Outreach.* ed. P. Bidgood [online]. Available at http://www.conkerstatistics.co.uk/iase/proceedings.php

Howden-Chapman, P., Matheson, A., Crane, J., Viggers, H. Cunningham, M., Blakely, T., Cunningham, C., Woodward, A., Saville-Smith, K., O'Dea, D., Kennedy, M., Baker, M.,

Waipara, N., Chapman, R., and Davie, G. (2007), "Effect of insulating existing houses on health inequality: cluster randomised study in the community," *British Medical Journal* [online], 334, 7591, 460. BMJ, DOI:10.1136/bmj.39070.573032.80.

Joiner, B. L. (1975), "Living Histograms," *International Statistical Review / Revue Internationale de Statistique*, 43, 3, 339-340.

Lindley, D. V., (2011), "Discussion on the Paper by Wild, Pfannkuch, Regan and Horton." *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 174, 2, 283–284. London: Royal Statistical Society.

Marchini, J. (2004), *Lecture 4: The Binomial Distribution.* [online]. Available at www.stats.ox.ac.uk/~marchini/teaching/L4/L4.notes.pdf

Maxwell, N. P. (1994), "A coin-flipping exercise to introduce the p-value," *Journal of Statistics Education,* [online], 2,1. Available at http//www.amstat.org/publications/jse/v2n1/maxwell.html.

McGillvray, H. (2011), Comment in question time 2011 International Association for Statistical Education (IASE) Satellite Conference: Statistics Education and Outreach, Malahide, Ireland.

Ministry of Education (2012a), *The New Zealand Curriculum Online. Mathematics and statistics curriculum achievement objectives* [online]. Available at http://nzcurriculum.tki.org.nz/Curriculum-documents/The-New-Zealand-Curriculum/Learning-areas/Mathematics-and-statistics/Mathematics-and-statistics-curriculum-achievement-objectives#.

Ministry of Education (2012b), *National Standards. Mathematics Standards. After two years at school,* [online]. Available at http://nzcurriculum.tki.org.nz/National-Standards/Mathematics-standards/The-standards/After-two-years.

Pagano, M. and Halvorsen, K. T. (1981), "An Algorithm for Finding the Exact Significance Levels of r × c Contingency Tables," *Journal of the American Statistical Association*, 76, 376, 931-934.

Pange, J. and Nikiforidou, Z. (2007), "The Notions of Chance and Probabilities in Preschoolers," *Early Childhood Education Journal.* 38, 4, 305-311. Springer.

Pant, M. (2006), "Unit 1. Introduction to Participatory Research" and "Unit 2. Conceptual Understanding Participatory Research" in *Certificate in International Perspective in Participatory Research*. New Delhi: Society for Participatory Research in Asia (PRIA).

Payne, B. "Nappy Changing Challenge and Classroom Olympics: Competitive and cooperative hands on data collection activities," in *Proceedings of the 2011 IASE Satellite Conference: Statistics Education and Outreach.* ed. P. Bidgood [online]. Available at http://www.conkerstatistics.co.uk/iase/proceedings.php

Pfannkuch, M., Forbes, S., Harraway, J., Budgett, S. and Wild, C. (2013), ""Bootstrapping" students' understanding of statistical *inference,"* [online], Research report for Teaching & Learning Initiative, Wellington, New Zealand. Available at http://www.tlri.org.nz/sites/default/files/projects/9295_summary%20report.pdf

Pratt, D. and Ainley, J. (2008), "Introducing the special issues on informal inferential reasoning," *Statistics Education Research Journal* 7, 2, 3-4.

Rossman, A. (2008), "Reasoning about informal statistical inference: one statisticians view," *Statistics Education Research Journal.* 7, 2, 5-19.

Russell, M., and Haney, W. (1997), Testing Writing on Computers: An experiment comparing student performance on tests conducted via computer and via paper and pencil," *Educational Policy Analysis Archives* [online], 5, 1. Available at http://epaa.asu.edu/epaa/v5n3.html

Schneiter, K. (2008), "Two Applets for teaching hypothesis testing," *Journal of Statistics Education,* [online] 16, 3. Available at http//www.amstat.org/publications/jse/v16n3/schneiter.html

Speed, T. (1986), "Questions, Answers and Statistics" in *ICOTS II, Proceedings. The second international conference on teaching statistics.* eds. R. Davidson and J. Swift, Canada: University of Victoria, pp.18-28.

Statistics New Zealand (2010), "Household Labour Force Survey: June 2008 Quarter – Media Release," [online]. Available at http://www.stats.govt.nz/browse_for_stats/income-and-work/employment_and_unemployment/household-labour-force-survey-info-releases.aspx

Statistics New Zealand (2011), *"Quick Stats About A Place." [online], Available at* http://www.stats.co.nz/Census/2006CensusHomePage/QuickStats/AboutAPlace/SnapShot.aspx

Statistics New Zealand, (2012a), "Household Labour Force Survey: June 2012 Quarter – Media Release," [online]. Available at http://www.stats.govt.nz/browse_for_stats/income-and-work/employment_and_unemployment/household-labour-force-survey-info-releases.aspx

Statistics New Zealand, (2012b), "Retail Trade Survey: June 2012 Quarter – Media Release," [online]. Available at *http://www.stats.govt.nz/*

Stirling, D. (2010), "Mastery tests to cope with mixed backgrounds in an introductory statistics course," in *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Ed. C. Reading, Voorburg, The Netherlands: International Statistical Institute. [online] Available at www.stat.auckland.ac.nz/~iase/publications.php.

Te Puni Kokiri (2002), *Survey of the Health of the Maori Language in 2001*. Wellington, New Zealand: Author.

Te Puni Kokiri (2007), *2006 Survey of the Health of the Maori Language. Final Report.* Wellington, New Zealand: Author.

Thompson, G. and Wilson, N. (1997), *Resource Document: A Brief History of Tobacco Control in New Zealand.* Commissioned by the Australasian Faculty of Public Health Medicine (NZ).

Unwin, A. (2003), "Mosaic Plots for Categorical Data," IBM workshop, Augsburg University, German.

Wang, H. and Harkess, C. (2007), *Senior Secondary Students' Achievement at Maori-Medium Schools 2004 – 2006 Fact Sheet*, Ministry of Education, Wellington, Demographic and Statistical Analysis Unit.

Wikipedia (2008), "Opinion polling for the New Zealand general election," [online]. Available at http://en.wikipedia.org/wiki/Opinion_polling_for_the_New_Zealand_general_election.

Wikipedia (2011), "Opinion polling for the New Zealand general election, 2011," [online]. Available at http://en.wikipedia.org/wiki/Opinion_polling_for_the_New_Zealand_general_election,_2011.

Wild, C and Pfannkuch, M. (1999), "Statistical Thinking in Empirical Enquiry.," *International Statistical Review / Revue Internationale de Statistique*, 67, 3, 223-248. International Statistical Institute (ISI).

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

Wild, C. J., and Seber, G. (2000), *Chance encounters: A first course in data analysis and inference*. New York: John Wiley & Sons.

Sharleen D. Forbes
School of Government, Victoria University of Wellington
PO Box 600
Wellington
New Zealand 6140
sharleen.forbes@vuw.ac.nz
