MAXINE PFANNKUCH

PROBABILITY AND STATISTICAL INFERENCE: HOW CAN TEACHERS ENABLE LEARNERS TO MAKE THE CONNECTION?

Chapter 11

> From the point of view of deductive logic that has shaped so much of statistical teaching in the past, probability is more basic than statistics: probability provides the chance models that describe the variability in observed data. From the point of view of the development of understanding, however, we believe that statistics is more basic than probability: whereas variability in data can be perceived directly, chance models can be perceived only after we have constructed them in our own minds. (Cobb & Moore, 1997, p. 820)

1. INTRODUCTION

Life is unpredictable and surrounded by seemingly random or chance events, or is it? It has become natural for people to observe that taller people tend to be heavier, that young men tend to have more car accidents, and that shares may return 3% per year in the long run. Inference or the drawing of a conclusion from data-based evidence abounds in the media. Crowded into a typical day's media articles are the results of polls, observational studies, and experiments such as: 58% of voters approve of the government's performance, 11-year-old children are getting heavier, or arthroscopic knee surgery is ineffective. In these scenarios inferences are respectively based on the following statistical elements: a random sample of 1000 people over the age of 18 and a quoted 3.2% margin of error, a sample of 11-year-olds taken in 1985 and in 2000 with a weight comparison between the two samples, and a comparison of a sample of patients randomly assigned fake surgery or real surgery. Such studies involve *statistical* inference, which attempts to draw a conclusion about a particular population from data-based evidence provided by a sample. Drawing inferences from data is now part of everyday life but it is a mystery as to why and how this type of reasoning arose less than 350 years ago (Davis & Hersh, 1986).

Graham A. Jones (ed.), Exploring probability in school: Challenges for teaching and learning, 267—294. ©2005

The foundations of probability and statistics were laid down as separate disciplines around 1660. Probabilistic thinking emerged in response to games of chance (Greer & Mukhopadhyay, this volume). Statistical thinking can be traced to John Graunt who, in 1662, used official statistics to draw inferences from data, such as estimating the population of London. These two quite different lines of thought took over 250 years to become connected. Stumbling blocks to incorporating probabilistic ideas into empirical data analyses appeared to be an inability not only to relate balls-in-urn-type problems to real world problems but also to believe that prediction was possible when there was a plethora of causes operating in the real world problem. Astronomers took a key conceptual step when they began to focus on the errors in their measurements rather than just the arithmetic mean of their measurements. Such recognition led to astronomy and geodesy using probability distributions, such as the normal distribution for measurement errors. Lightner (1991, p. 628) described this realization as a transition phase because "many concepts from probability could not be separated from statistics, for statisticians must consider probabilistic models to infer properties from observed data." It was not until the end of the 19th century, however, that the astronomers' error curve was reconceptualized as a distribution governing variation in social data such as heights of people. During the first half of the 20th century statistical inference based on probability became integrated into the discipline.

The recognition that mathematical probability models could be used to model and predict group (e.g., human group) behavior resulted in a shift in thinking that incorporated a nondeterministic view of reality. Historically, there were huge conceptual hurdles to overcome in using probability models to draw inferences from data; therefore, the difficulty of teaching inferential reasoning should not be underestimated.

Research on students' informal and formal inferential reasoning would suggest that there are huge gaps in current knowledge about how best to enable learners to make the connection between probability and statistical inference. The integration of statistical data analysis with theoretical probabilistic distributions and the assumptions underlying those models present a real conundrum in teaching.

Biehler (2001) argued that there was a four-stage development process for refining students' thinking towards formal inference. In the context of an example involving the comparison of two boxplots he described the stages and the person-roles as: fine tuning the comparison (the EDA methods expert); widening and exploiting the context by bringing in more variables (the subject matter explorer); generalization (the critical theory builder); and chance critique: Can group differences be "due to chance"? (the inference statistician). Whatever the developmental process should be, there is a need to build a pedagogical framework in order to develop students' inferential reasoning. Such a framework would give teachers a sense of the overall aims and purposes of statistical inference and the statistical reasoning processes that need to be developed when they teach the prescribed curriculum content. Without attention to the complexity of informal inference and to the provision of a teaching pathway towards formal inference, statistical and probabilistic inferential reasoning will continue to elude most students.

This chapter considers a possible pathway to formal inference by first drawing on, as an illustration, a case study that involved students in drawing informal inferences from the comparison of boxplots. Second, ways that students could be helped towards formal inference are suggested, and finally two possible pathways to formal inference, theoretical or simulation, are discussed.

2. INFORMAL INFERENCE

Before students are introduced to confirmatory or formal inference methods to decide whether the patterns they see in data are real or random, they are usually presented with situations that require informal inference. Research on students' informal inference from comparison of data plots is relatively recent. Biehler (1997) analyzed a transcript and videotape of some Grade 12 students' methods of handling multivariate data. From the perspective of how a statistical expert would handle the data he identified a number of problem areas for teaching data analysis. In particular for the comparison of boxplots he pointed out the difficulty of drawing conclusions, even for experts, when faced with a variety of patterns and when encountering differences in medians, ranges, and interquartile ranges each of which may support differing conclusions. He acknowledged the difficulty of verbally describing and interpreting graphs, and reported that the language used by both teachers and students was inadequate.

Konold, Pollatsek, Well, and Gagnon's (1997, p. 165) analysis of the same Grade 12 students that Biehler had used, hypothesized that the students had not made "the transition from thinking about and comparing properties of individual cases or collections of homogeneous cases to thinking about and comparing group properties". The desired thinking was described as a propensity perspective, the development of which the authors were not, at that time, prepared to prescribe. McClain, Cobb, and Gravemeijer (2000), however, believed that their instructional experiments, designed to focus seventh-grade students' argumentation on how the data were distributed, developed students' ability to reason about group propensities. The students' argumentation, for example, suggested that 75% of the observations for treatment X were greater than 75% of the observations for treatment Y and therefore treatment X would be recommended. Argumentation issues such as sample size and sampling variability were not broached and would not be expected at this level. Ability to take into account the sample size when drawing inferences from data is described by Watson (2001) as a higher order skill.

In fact, Konold and Pollatsek (2002) recommended that the early teaching of statistics should focus on informal methods of data analysis. They envisaged that the focus should be on why the data are collected and explored and what one learns from the data. Their idea of a data detective approach to data analysis fits with that of Pfannkuch, Rubick, and Yoon (2002), who believe students should approach data analysis in the thinking roles of hypothesis generator, discoverer, and corroborator. In other words, statistical exploratory data analysis should largely be kept separate from probability, with only informal quantifications of variability to denote a propensity or a spread of one sample distribution compared to another. It should be noted that Shaughnessy (2003) advocates that the teaching of probability should always be connected to a statistical approach. Furthermore, he suggests that previous recommendations to start with a probability problem and then gather data should perhaps be the other way around. That is, "statistics should motivate the probability questions" (p. 223).

A Case Study

The following case study of some Grade 10 (15-year-old) students' attempts at informal inference is used to illustrate how and why a proposed framework for transitioning students towards formal inference needs to be formulated.

Background

In 2002 a new approach to national assessment in New Zealand was introduced at Grade 10. Instead of one final external examination in mathematics, one third of the course is internally assessed, with external moderation, and the rest is an external examination (New Zealand Qualifications Authority, 2001). Statistics is internally assessed and students are given data sets to investigate. The level of statistical thinking required at Grade 10 with this new internal assessment, compared to the previous external assessment that largely asked students to read and interpret graphs and calculate measures of central tendency, produced real challenges for teachers and students. These challenges set the scene for the case-study investigation.

Research method

Based on the ideas of Gravemeijer (1998) and Skovsmose and Borba (2000), a developmental research method was used (Pfannkuch & Horring, 2004). The school selected, which draws on students from low socioeconomic backgrounds, is culturally diverse, and has teachers interested in improving their statistics teaching. A workshop, which focused on communicating the nature of statistical thinking to the teachers (Wild & Pfannkuch, 1999), was conducted by the author. After the workshop the self-selected case-study teacher and another teacher were interviewed to identify problematic areas in their statistics teaching (Pfannkuch & Wild, 2003). The teachers then wrote a new 4-week statistics unit. Although all Grade 10 teachers implemented the new teaching unit, research data were collected mainly from the case-study classroom. These data were videotapes of 15 lessons, student bookwork, student responses to the assessment tasks, student questionnaires, and the teacher's weekly audio-taped reflections on the teaching of the unit. Teacher and author observations as well as the student responses on the questionnaires identified informal inference as a problematic area. Therefore the first analysis of the assessment task data focused on how students drew inferences from data. The results of the analysis led to a consultation group of five statisticians being formed to debate and discuss possible ways to progress.

The assessment task

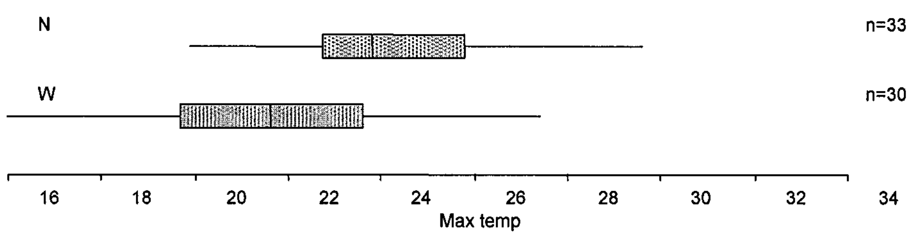

Figure 1. Comparison of Napier (N) and Wellington (W) maximum temperatures in the summer

The students were given a table of data showing the maximum temperatures of two cities, Napier and Wellington, which were presented to the students as if they were two independent samples. At a much more sophisticated level, analyses taking into account pairing would be done. A story involving a decision about where to go for a summer holiday was communicated to the students. Students were required to pose a question (e.g., Which city has the higher maximum temperatures in summer?), analyze the data, draw a conclusion, justify the conclusion with three supporting statements, and evaluate the statistical process. All students chose to analyze the data by calculating the five summary statistics and then drawing boxplots by hand. Figure 1 shows the boxplots drawn electronically. Note that Grade 10 students are not expected to identify outliers so the whiskers were drawn to the minimum and maximum observations.

Results

The analysis of the student assessment responses was grounded in the hierarchical performance levels associated with the SOLO taxonomy (Biggs & Collis, 1982). Based on the student responses, four categories of justifications for their conclusions were identified: comparison of equivalent summary statistics; comparison of nonequivalent summary statistics; comparison of variability; and comparison of distributions. Within these categories hierarchies of responses according to the SOLO taxonomy were identified and qualitatively described: *no response*; *prestructural* – irrelevant information; *unistructural* – some relevant information but non-discriminating; *multistructural* – some relevant information with some discrimination; and *relational* – information communicated is relevant to the question and is discriminating. After the qualitative descriptors for each category and each level within a category were written by the author, the author and another person independently coded all responses. A consensus was then reached on the final codes for each student response. The details of the student responses are recorded in Table 1 and some examples of student responses are given in Figure 2.

Conclusion responses ranged from nonuse of comparison language to comparisons that suggested statistical tendency. The analysis of the justification statements for the conclusion revealed that students compared features of the boxplots in a nondiscriminating manner (unistructural responses). Students did not tend to explain how their analysis supported their conclusion and was appropriate in relation to the question (relational responses).

	Conclusion	Comparing equivalent stats	Comparing non-equivalent stats	Comparing variability	Comparing distributions
No response	2	3	12	9	21
Prestructural	0	2	3	1	0
Unistructural	11	8	7	15	9
Multistructural	11	7	4	5	0
Relational	6	10	4	0	0
Total number of students	30	30	30	30	30

Table 1. Details of student responses when comparing boxplots

Comparing equivalent summary statistics (27/30) and comparing the variability with a statement about the ranges that was not relevant to the question posed (16/30 classified as pre- or unistructural responses), were prevalent in student responses. Eighteen students attempted comparison of nonequivalent summary statistics (see Fig. 2). There was no attempt at comparing the difference in medians in relation to the variability (e.g., Is the difference between the medians quite large or small compared to the variability in maximum temperatures between Napier and Wellington?) and little attempt at comparing the shape of the distributions.

A qualitative analysis of the learning experiences provided, using the videotape and student bookwork data, suggested that students had learning opportunities that only compared features of the data. For example, the teacher's only worked example with the class was one where the question generated was: "What are the differences/similarities between male and female exam scores?" Hence only features, such as "the IQR for the male data is smaller than for female and therefore less spread," of the sample distributions were compared in class.

The last part of the assessment task required students to evaluate the statistical process with three separate statements. Twenty students said that more data should be made available before making a decision. For thirteen students, however, a typical comment was:

> Firstly Wellington only has 30 temperatures where as Napier has 33. Giving Napier an unfair advantage. For this to be a fair test there needs to be exactly the same number of temperatures. Those 3 extra temperatures have affected the result.

Even though students had compared data sets of unequal size in class they were not asked to raise concerns about the comparison and hence their belief that data sets should have equal sample sizes was not uncovered.

Conclusion
- Napier has the highest temperature (U).
- Napier is warmer than Wellington (M).
- Napier tends to be warmer than Wellington (R).

Comparing equivalent summary statistics
- Also the fact that the statistics for Napier are higher than Wellington (except the interquartile range) (U).
- Napier has a higher median than Wellington. Napier has a highest temperature of 33.1°C but Wellington's highest is 27.4°C (M).
- This is shown in the median, with Napier's median being several degrees higher than Wellington's (R).

Comparing nonequivalent summary statistics
- Napier's median temperature is higher than Wellington's upper quartile (U).
- Also because the median of Napier's temperature is higher than three quarters of Wellington's temperatures which suggest that half of Napier's temperatures are warmer than three quarters of Wellington's (R).

Comparing variability
- Napier has a larger range of data compared to Wellington (U).
- The box-and-whisker plot also shows that Napier has a wider range of temperatures, and that many of the temperatures are grouped between 22.75°C and 23.8°C, while in Wellington the temperatures are more evenly spread (M).

Comparing distributions
- The box plot for Wellington is drawn lower than Napier's (U).

Figure 2. Examples of student responses with SOLO level indicated

Discussion

Hypotheses were generated by the author and five statisticians as to why drawing a conclusion and justifying it were problematic when comparing data plots. One hypothesis was that school and introductory textbooks and therefore teaching tended to compare only features of boxplots and not to draw a conclusion, since significance testing and confidence intervals are introduced at a later stage (Wild & Seber, 2000). Other hypotheses proposed were as follows: the assessment demands were beyond the capabilities of Grade 10 students, 'informal inference' techniques were not established or recognized within the statistics discipline implying that the assessment expectations were unrealistic, the curriculum did not provide a teaching pathway to build students' concepts of formal inference or provide learning experiences for the transition between informal and formal inferential thinking.

Informal inference could have been presented to the students by giving clear-cut examples and limiting them to comparing data sets of similar spreads and samples of size 30. This was not what the teachers wanted; they considered it was the inherent messiness of data, the absence of a clear decision, and the positing of possible contextual explanations, which made data comparison interesting. If informal inference was to be taught there would need to be more awareness among teachers of the formal inference ideas underpinning comparison of data plots.

In thinking of the needs of informal inference for the comparison of data plots, the author and statistician group determined that before drawing a conclusion there were four basic aspects to attend to in order to understand the concepts behind significance tests, confidence intervals, *P*-values and so forth. These were identified in the following way: comparisons of centers; considering the differences in the centers relative to the variability in the samples; checking the distribution of the data (normality assumptions, outliers, clusters); and the sample size effect. In cognizance of these conceptual underpinnings for formal inference and of the student responses, a pedagogical framework *towards formal inference* is beginning to be developed. This framework, based on the assumption that formal inference notions should begin to be developed by Grade 10, continues to be under debate. It is a framework for making teachers aware of the reasoning that students need to experience and develop for inference, namely:

1. Reasoning with measures of center
2. Distributional reasoning
3. Sampling reasoning
4. Drawing an acceptable conclusion based on informal inference.

Underlying this reasoning is a fundamental statistical thinking element, *consideration of variation* (Moore, 1990; Wild & Pfannkuch, 1999). It is this consideration of variation that is closely allied to developing students' probabilistic reasoning.

3. TOWARDS FORMAL INFERENCE

During Grades 9 to 12 the connections between probability and statistics should gradually be developed and informally introduced to students. In particular, when using the pedagogical framework *towards formal inference*, attention needs to be drawn to a number of key principles: probability can be used to quantify variability, data can be modeled by probability distributions despite the multitude of causes operating, confidence limits or boundaries exist, and samples are drawn from populations. Relations between data analysis and probability have to be consciously developed in teaching (Biehler, 1994). Some ways of addressing the four components of the pedagogical framework in teaching are now discussed.

Reasoning with Measures of Center

Wild and Pfannkuch (1999, p. 240) said that "the biggest contribution of statistics is the isolation and modeling of 'signal' in the presence of 'noise'". If the comparison of boxplots is considered then the medians are the signal and the variability within and between the boxplots is the noise. Two-thirds of the case-study students did not acknowledge that the comparison of the medians was the salient feature of the statistical comparison. Such a finding resonates with Konold and Pollatsek's (2002, p. 273) research which found that students failed "to interpret an average of a data set as saying something about the entire distribution of values". They believed that statistical reasoning would elude students unless they understood that the comparison of averages is the statistical method for determining whether there is a difference between two sets of data.

From the learning experiences observed in the case-study classroom, the students would be unlikely to know why the comparison of centers should be the focus of their reasoning. Konold and Pollatsek (2002) suggest that the central idea of searching for a signal amongst the noise has not been the focus of teaching and hence students have not developed this notion. The learning experiences that they suggest involve students appreciating causal-type variability in a process, its inherent probabilistic nature, and the consequent building of a mound-shaped distribution. Biehler (1994) contrasts the explaining and describing of variation by causal and other factors and consideration of probability models as two cultures of thinking, namely EDA thinking and probabilistic thinking. He believes that the connections and interfaces between the two modes of thinking are problematic. He concurs however with Konold and Pollatsek (2002) and Joiner (1994) that the Galton board is a useful basis for teaching such connections. The Galton board has a funnel containing beads, rows of pins that simulate factors acting on a process, and a series of slots into which the beads fall. The process of the beads falling through the funnel (signal), then bouncing through the pins (noise), and finally forming a mound shaped distribution in the slots physically embodies the signal and the noise.

Learning experiences based on these ideas could help students to reason with and search for the signal or measure of center amongst the noise.

Joiner (1994) advocates that students should collect and plot data such as the arrival time of their school bus each day over an extended period in a time series plot (Fig. 3). From such a graph the underlying signal (mean) and the noise can be visualized, and boundaries that capture most of the variation can be informally imposed. The time-series graph also allows students to visualize and discuss ideas of randomness whereas a mound-shaped graph does not. From a quality management perspective the strategy for interpreting such a graph is to differentiate between special-cause and common-cause (or chance) variation. Since it is natural for students to look for causes these intuitions can be built into teaching.

The first strategy is to identify special-cause variation, which is usually outside the upper and lower limits. These data come from outside the usual process, which in this example can be identified as the time when the bus was very late. Possible reasons for such an unusual occurrence could be a bus breakdown, or a passenger becoming ill, or a driver not turning up for work. These reasons are not part of the normal expectations or occurrences for the driving of a bus. Special-cause variation can contribute either a small or large amount to total variation and typically has a much bigger impact on variation than any common-cause variation.

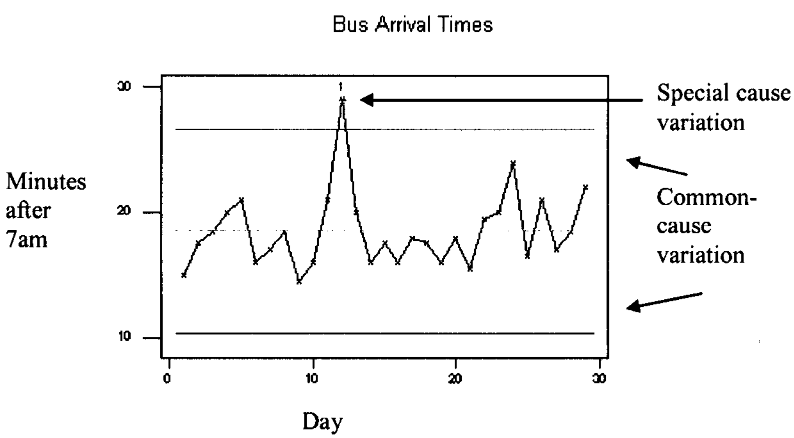

Figure 3. Student investigation of her bus arrival time (adapted from Joiner, 1994)

Common-cause (or chance) variation is confined between the upper and lower limit times. For the bus example some common cause factors could include normal day-to-day variation: traffic density, number of people catching the bus, green light run, weather, and other factors of which one is unaware. These factors are present all the time. Individually they have a small effect but collectively they can add up. For example, if the weather was bad, the traffic congested, and many people were catching the bus, then collectively these effects might compound and the bus may well be much later than usual. This type of discussion about the data should give students the notion of variation not being attributable to one cause but a multitude of causes that are modeled as random or chance variation. Such variation cannot be explained. This investigation would highlight thinking about variation in terms of realizing that variation happens, that some of it can be explained, but the rest cannot. This random variation is described mathematically by probability. According to Wild and Pfannkuch (1999, p. 242), "special-cause versus common-cause variation is a distinction which is useful when looking for causes, whereas explained versus unexplained variation is a distinction which is useful when exploring data and building a model for them."

The same bus arrival data (Fig. 3) can be used to construct a mound-shaped graph (Fig. 4), a graph which is the result of a random process that has no detectable pattern. There should be recognition that even though an individual event (arrival time) cannot be predicted, the group as a whole obeys some law of stability and hence predictions can be made about the behavior of the group. It would seem that students should first construct a series graph to visualize and experience the random variation and signal, and second, construct a mound-shaped graph in which the signal and noise are represented in a different perhaps nonintuitive way. Such learning experiences including those suggested by Konold and Pollatsek (2002), as well as other similar approaches for probability experiments, could lead students to a nondeterministic or probabilistic view of reality. Probabilistic thinking helps separate the reality or signal from the background noise. The link between explaining and describing variation by causal and other factors and modeling the variation by probability distributions (Biehler, 1994) is crucial in relating probability and data analysis.

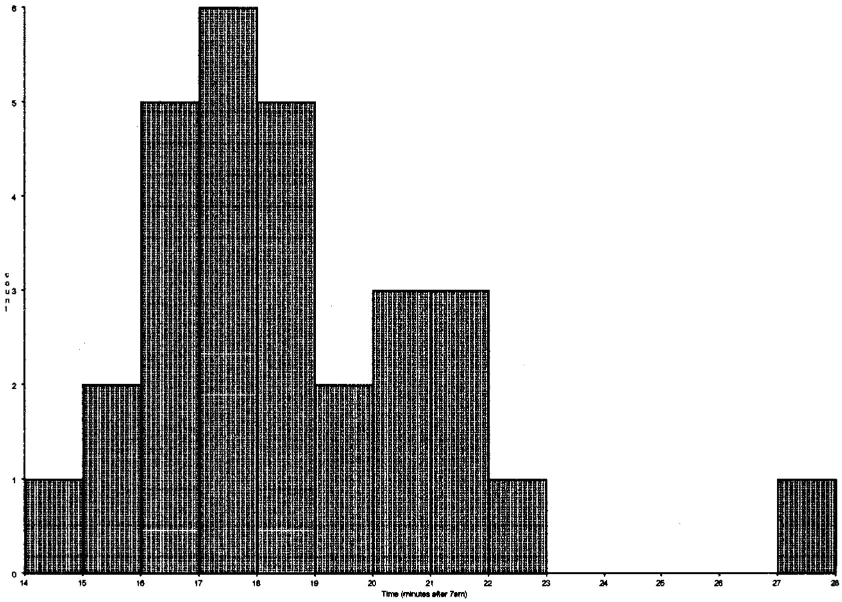

Figure 4. Student investigation of her bus arrival time distribution

Distributional Reasoning

When comparing variability between Napier and Wellington temperatures, half of the case-study students made a comparison of the ranges, which was not relevant to the question they posed. Formal inference requires considering the differences in the centers relative to the variability in the samples, which presupposes an understanding of standard deviation or confidence intervals or significance. Taking the sample size into account a statistician might, but is unlikely to, informally infer by mentally intuiting confidence intervals for the true population means and visualizing whether there might be an overlap. This would be an impossible inference for a Grade 10 student with no experience of confidence intervals. The students, however, could look at variability within a data set and between data sets. The focus in teaching could be on describing, interpreting, and comparing the variability in the data sets rather than attempting to determine whether event A is "greater" than event B. For example, Wellington has a fairly symmetrical distribution whereas Napier is less symmetrical with some bunching – greater density – between 22.75 (LQ) and 23.8 (MED) and a greater spread between 25.5 (UQ) and 33.1 (MAX). It should be noted, however, that the students would be describing sample distributions rather than population distributions and the features they note may well be due to "chance" (see Fig. 5).

Recent research (e.g., Konold et al., 2002; Watson, 2002) has focused on distributional reasoning and the importance of students describing the "clumps and bumps" and attending to the variability within a distribution, which by its nature is both probabilistic and statistical. Experiences that involve looking at distributions of data are prerequisites to experiencing the behavior of random events and the probability distributions that describe them. To build up concepts about distribution as well as confidence intervals, the connections between probability and statistics can be reinforced through consideration of the range of "likely" outcomes in repeated probability experiments. As Shaughnessy (2003) notes, "Confidence intervals model the variability of the likely point values from repeated probability experiments. The concepts of sample space and variability are closely connected" (p. 223). According to Scheaffer, Watkins and Landwehr (1998, p. 17) "probability questions should require students to observe the entire distribution rather than just the height of one bar." Links could also be made between the distributions and variability in probability experiments and the distributions and variability present in social data.

In particular students should not continue to believe that comparing a feature such as "50% of Napier's temperatures are higher than 75% of Wellington's temperatures" is evidence for a real difference, rather that it may be a noteworthy feature to describe. It is important also that students consider that the difference may have resulted from chance. According to Moore (1990) and Konold (1994, p. 206) "students do not spontaneously raise this possibility." The term "chance" should not be lightly overlooked in teaching, as students may understand the term in dice problems but may not for real problems where causes are known (Wild & Pfannkuch, 1999). What students should be building up is the concept that they have sample data and that if they took other samples they would obtain different plots.

Sampling Reasoning

Statistical inference reasons from the sample to the population, a notion that is alien to most students according to Scheaffer et al. (1998), whereas probability reasons from the population to the sample. Taking samples from a hypothetical population and recognizing the importance of sample size are major problems associated with informal inference. There are many strands to building up concepts about sample size effect. Based on the case-study some matters that need attention are the following: comparison of sample data of unequal sample size; the notion of a sample; the sample and its relationship to the population; the size of the sample and its relationship to the population; and small sample versus large sample variability. A repertoire of teaching and learning possibilities needs to be considered to build up these concepts (Watson, 2004). Biehler (1994, p.15) considers that "the production of random samples from populations and the randomization in experiments should be an intermediate step that consolidates the conceptual shift from data analysis to inference." Whereas from the perspective of building up students' conceptions of sample and sampling, Saldanha and Thompson (2002) suggest reinforcing schema that interrelate repeated random selection, variability among outcomes, and distribution.

For the case-study students it was necessary first to overcome the belief that the data sets must be of the same size. Using Curcio's (1987) hierarchical model for interpreting graphs, the author's observation, corroborated by the teachers, was that the students had experience of reading the data, less experience at reading between the data, and little experience of reading beyond the data. If these students had some experience of inferring "missing data" from a data set they may have predicted that the missing summer temperatures were likely to be within the interquartile range or at least within the range. Such informal probabilistic notions are essential in building up ideas about likelihood and confidence intervals. The problems of missing data are well-known in statistics and students could be given opportunities to impute values for observations and to analyze data with and without the imputations. Specific attention could be drawn to students' beliefs and to whether their conclusions would change with unequal sample sizes.

According to Watson (2004), despite curriculum statements referring to sampling and making inferences about populations, there has been a lack of attention to reasoning about samples and sampling in schools. She hypothesizes this situation may have arisen because "developing reasoning related to sampling may be associated with developing literacy and social reasoning skills rather than developing numeracy skills." Her research suggests that students have inadequate foundations for making inferences. Students know variation exists in a population and their dilemma is how to deal with this variation when taking a sample. The idea of taking random samples, a chance process that ensures appropriate variation, appeared to be a sophisticated notion. Students struggled with conflicting ideas such as suggesting that the sampling method should ensure representativeness, include all the population, and be fair about who should be included. She suggests class debates could be used to convince students of the need to take random samples and suggests that the range of views reported by Metz (1999) would be a good starting point. Random sampling is critical to inference. The reasoning underlying random sampling is not based primarily on calculations but on an appreciation of the role of probability and variation in the selection of samples that will be used to make inferences about a population.

Saldanha and Thompson (2001) reported on a teaching experiment in which students drew random samples from unknown populations of objects such as candies in a sack. In order to make a decision about whether two students' results were unusual all the students engaged in determining a criterion for deciding when two distributions were alike. According to Saldanha and Thompson the critical shift in students' thinking was turning away from thinking about single samples to considering the proportions of the collections of samples that were similar to the unusual distribution. Such informal significance testing using hand-drawn samples as well as computer simulations is important for starting to understand that the cut-off point for making a decision is arbitrary but the uncertainty can be quantified by probability. Similar scenarios involving balls-in-urn type problems as well as real-world problems that can be simulated could become part of the teaching repertoire so that students experience making their own decisions about whether an effect is real or not. Students can also decide on the size of the sample to take and through such experiences "build an understanding of a 'reasonable size' for a representative sample ... and form a link between reliability and sample size" (Watson & Moritz, 2000, p. 133).

The problem of informal inference is compounded by sample size and variability being interconnected. Watson and Moritz (2000) suggested explicit discussion with students would be profitable. Aspinwall and Tarr (2001) indicated in their research, with balls-in-urn type problems, that most middle school students could obtain an appreciation of the role of sample size in uncovering the parent distribution. Focused instruction on drawing students' attention to the number of trials and the outcomes of the probability simulations was part of the teaching design. In another study (Stohl & Tarr, 2002), two students, using probability simulation tools and carefully designed teaching tasks, were facilitated to establish a link or bidirectional relationship between empirical and theoretical probability, to understand the role of sample size, and to make inferences from probability experiments.

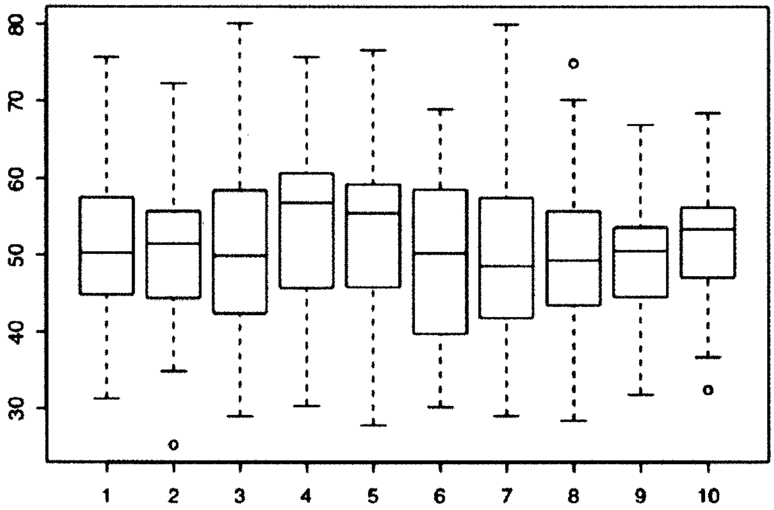

Figure 5. Randomly generated samples of size 30 from a Normal distribution with mean 50 and standard deviation 10

Hand and computer simulations, such as taking random samples of the same and different size, as well as small and large size, from a population, could also be part of students' learning. Such simulations might enable students to "see" the variability of the sample mean, the variability within and between samples, and allow comparison of the variability of the samples with the variability of the population (e.g., Fig. 5). Context-free simulations such as shown in Figure 5 might not advance students' conceptions of the sample size effect but hypothetical situations grounded within the context of a problem (e.g., If they took another summer's temperatures would they get the same graphs? How large a sample should they take to be fairly certain that it was representative of the population?) might start to induct students into some formal inference ideas. Formal inference is predicated on the idea of "taking random samples" and independent observations, which in such contextual situations may not be the case but the model assumptions can be discussed at a later stage.

Drawing a Conclusion

Consider the case-study example: If there were no overlap between the boxplots, statisticians would not carry out a formal test for no difference between the means. Such a test may be required, however, when plots are considered to be overlapping. Simulations could be used to overcome students' belief that the statement "50% of Napier's temperatures are higher than 75% of Wellington's temperatures" is evidence for a real difference as well as the belief that a sample size of 30 is large enough. Students' attention could be drawn, for example, to noticing that some randomly generated plots of sample size 30 from the same population distribution give rise to the above statement: see boxplots 2 and 4 or 9 and 10 in Figure 5. The simulations should generate boxplots and histograms, as these are the types of graphs from which the students are required to make informal inferences.

For drawing a conclusion the statistician group suggested that Grade 10 students could "look at the plots" and compare the centers, spreads, and anything else that is noteworthy. After comparing and describing features, students could then draw an informal inference: for example, "the sample data suggest that Napier has higher maximum temperatures on average in summer than Wellington." The words "sample", "suggest", and "on average" were used to reinforce statistical inference ideas. By Grade 11 students could be referring to the underlying population when drawing conclusions. The question of whether the students are drawing a valid conclusion can be addressed in another pedagogical framework that focuses on the evaluation of the statistical process. Questions can be raised such as: "Does this conclusion make sense in terms of what I know about the real world? Is there an alternative explanation?"

Many rich learning experiences, particularly in the above four components of the pedagogical framework, are necessary to prepare students for formal inference. All areas require at a foundational level *consideration of variation,* which by its very nature is linked to probabilistic understandings.

4. FORMAL INFERENCE

Fundamental to statistical inference is the recognition that sample data can be used to make predictions and decisions about the underlying population and that the sample selected is just one of many samples that could be drawn from the population. Underpinning formal inference methods are understandings of sampling distribution, random sampling, and distribution of the mean differences as well as recognition that the comparison of measures of centers is central to the argument. Recent research suggests that better teaching methods are needed to improve students' conceptual understanding of sampling in relation to statistical inference (Watson, 2004). Research on formal inference is limited. Educators, however, are attempting to find new ways, using computer-based simulation approaches, to improve students' conceptual understanding of statistical inference and the probability distribution models that are used.

Amongst statisticians (e.g., Cobb & Moore, 1997) there is general agreement that significance tests are overused and that the size of the effect is usually more important than how statistically significant it is. Therefore confidence statements should be introduced before significance testing. Scheaffer et al. (1998, p. 23) believe that the approach to confidence intervals should be through simulation so that "students can begin 'to develop some feel' for reasonable values of population parameters" before formalization.

Classical Approach

Traditionally the approach to inference is a probability theory-based explanation couched in mathematical language. The rationale, however, for this approach is obscure to most students. Consider the example used in the case study and the process of reasoning that older students would carry out for a significance test. First the students would establish a null and a one-sided or two-sided alternative hypothesis for the underlying population means. On the assumption that the two samples were randomly selected and taken independently from two normally distributed populations, with underlying means μ_N and μ_W and with unknown standard deviations, a significance test would be conducted. Assuming the null hypothesis was true a standard statistical package calculates the test statistic, confidence intervals, standard error, *P*-value, and degrees of freedom. The test statistic has a probability distribution that can be approximated by the *t*-distribution with n_N+n_W-2 degrees of freedom, where n_N and n_W are sizes of the samples, $\bar{x}_N$ and $\bar{x}_W$ are the means of the samples, and s_N and s_W are the sample standard deviations. On the basis of the *P*-value the students would assess the strength of evidence against the null hypothesis and then conclude that there was no/weak/strong evidence of difference in mean maximum temperatures between Napier and Wellington. Or more traditionally, decide whether or not to reject the null hypothesis. Inference in this case might be applied to data that are not the product of random sampling. There is an assumption that a probability model does govern the data production. And of course the degree of uncertainty includes only the chance variation, it says nothing about other potential sources of error.

According to Moore (1990, p.134), in the classical approach "the mechanics of stating a hypothesis, calculating a test statistic" and finding a *P*-value "conceal the reasoning of significance tests". He believes the reasoning is difficult and significance tests need not be in the school curriculum. Not only is the reasoning difficult but, in addition, the myriad of underlying concepts that are theoretically addressed in the lead up to statistical significance testing remain largely elusive to most students.

Simulation and Classical Approach

With more technology becoming available, many educators who recognized that the theoretical development was deficient, started to use simulations, particularly for the sampling distribution and the Central Limit Theorem. Despite the promotion of the use of such simulations in instruction, delMas, Garfield, and Chance (1999) concluded that there was no substantial evidence that simulations actually improved students' conceptual understanding of the sampling distribution. Lipson (2002), for example, focused her research on elucidating tertiary students conceptions on sampling distributions and hypothesis testing. Based on experience and other research evdience that the concept of sampling distribution was poorly understood and that an empirical view of sampling distribution was an essential component of students' schemata, Lipson (2002) exposed mature-age students to a learning strategy that involved dynamic computer simulations of the sampling process linked to the formation of a sampling distribution. A common confusion among students initially was the difference between the distribution associated with the sample and the sampling distribution. Student concept maps revealed that the sampling software helped in elucidating some aspects of sampling distributions, but failed to link the empirical and theoretical representations of the sampling distribution and to link the sampling distribution to hypothesis testing and estimation. She concluded that instructional improvements in software for the development of the concept of sampling distribution were needed.

In a recent small study Lipson, Kokonis, and Francis (2003) devised a computer simulation session to support the development of students' conceptual understanding of the role of the sampling distribution in hypothesis testing. They reported that students' conceptual understanding progressed through four developmental stages: (a) recognition (the sampling distribution summarizes repeated samples from a hypothesized population, the sample statistic is variable), (b) integration (locating the observed sample on the hypothesized sampling distribution, the concept of a single population), (c) contradiction (recognizing an inconsistency between observed sample and hypothesized population, one observed sample but a range of possible populations from which it may have been drawn, extending the concept of variability to the population), and (d) explanation (possible statistical explanations for contradiction, decisions based on probabilities). A stumbling block for students appeared to be that they looked for a contextual explanation rather than a statistical explanation, even when they "acknowledged the low probability of the sample coming from the hypothesized population" (p. 7). They concluded that current software supported the recognition stage only and that students need considerable support for the other developmental stages. They also suggested that "students need to have a lot more experience in thinking about the kinds of samples that one could expect to arise from the sampling process" (p. 9).

DelMas et al. (1999) also sought to improve students' conceptual understanding of sampling distribution. They created their own software and course materials, which, after several iterations, are becoming more effective at challenging students' understanding of the sampling distribution. They found that "good software and clear directions that point students to important features will not ensure understanding" (p. 8). Rather, course activities are needed to challenge each student's misconceptions. Better results were obtained when the activities were structured "to help students evaluate the difference between their own beliefs about chance events and the actual empirical results" (p. 8).

Simulation Approach

Although there is beginning to be some success at improving students' understanding of sampling distribution and its relationship to significance testing, this is only one of many critical steps in developing statistical inference expertise. Jones, Lipson, and Phillips (1994) argue that attempts at using an empirical approach, such as exposing students to computer simulations to build up the concept of a sampling distribution, have largely been ineffective. They conjecture that a reason for this problem is that students have difficulty in integrating "their empirical experience of the sampling distribution with the theoretical model of the sampling distribution that is used in classical inference" (p. 257). They argue that the theory-based approach is inaccessible to today's cohort of students taking introductory statistics. They advocate, as do other educators (e.g., Scheaffer, 1992; Konold, 1994; Biehler, 2001) that inference should be dealt with entirely from an empirical perspective. Scheaffer (1992, p. 80) believes that students who are taught through simulation methods will "understand how statistical decisions are made" and more importantly realize that classical procedures "require an underpinning of randomness". Some computer packages designed for students such as *DataScope* (Konold & Miller, 1995) and *Fathom* (Key Curriculum Press Technologies, 2000) enable this approach.

The resampling approach, initially proposed by Julian Simon in the late 1960s (Konold, 1994), has the potential to make strong connections between probability and data analysis. It elucidates how probability provides a theoretical structure for statistical inference, as it is based on the notion of considering what would happen if the method was repeated many times. Consider the case-study example where the observed median difference in maximum temperatures is 2.2. The student can ask the question: "Although I know that my particular samples for Napier and Wellington maximum temperatures have these particular medians and spreads I know that if I repeat this study with the same sample sizes I will get different values. So, is this difference I see between Napier and Wellington maximum temperatures due to 'chance' (random or sampling variation) or is there a real difference?" The resampling method takes all the maximum temperature values and randomly reassigns them to Napier and Wellington. Given that the values have been randomly assigned then any difference between the medians of Napier and Wellington is due to "chance". Students can then look at some of the plots generated by this procedure and state: "If the difference were due to chance (or random or sampling variation) then I could obtain graphs like the following (Fig. 6)."

The computer can then be instructed to repeat this procedure, say 1000 times, each time computing the difference in medians. A histogram of the empirical distribution of the differences in medians (Fig. 7) can then be displayed and a one-sided *P*-value can be estimated. In this case the student can state: "Suppose that the difference is due to 'chance' (random variation) how often will I see this difference of 2.2 or larger in the medians? This difference or larger occurs less than once in 1000 through 'chance' (random variation) only. Therefore there is strong evidence of a real difference in median maximum temperatures between Napier and Wellington."

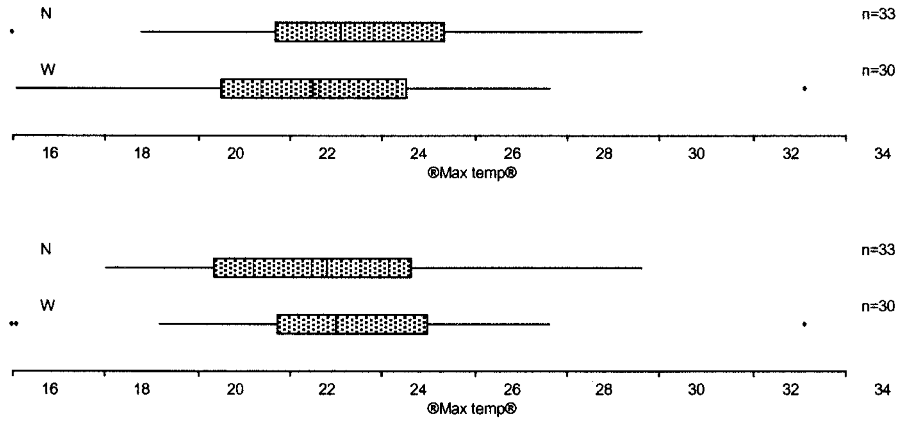

Figure 6. Maximum summer temperatures randomly assigned to Napier (N) and Wellington (W)

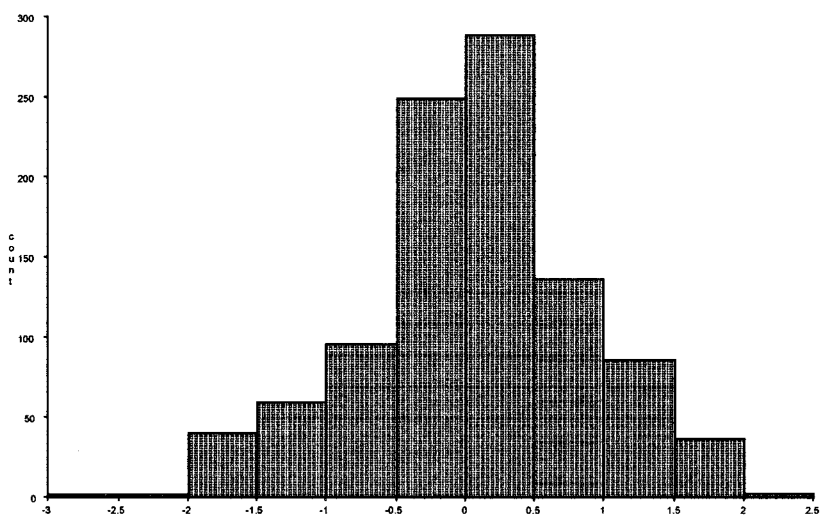

Figure 7. Empirical sampling distribution of the difference in medians for 1000 resamplings

Classical statistical inference presupposes a decision-making process is in operation. The logic of P-values, however, is based on the degree to which the outcome was surprising, which is more about assessing the strength of the evidence for the populations under consideration (Cobb & Moore, 1997). The resampling approach would seem to be more in tune with such a conceptualization of a P-value.

Konold (1994, p. 210) used the resampling approach in one course with high school students. He reported, "at a deeper level, many students after instruction using resampling appear unaware of the fundamental nature of probability and data analysis." He believed, however, that the resampling approach in instruction should not be discarded but that it should be trialled over a series of courses to determine whether conceptual understanding was possible. Furthermore, Simon, Atkinson, and Shevokas (1976) reported that students who used the resampling approach consistently outscored the students using the traditional approach. Research is limited at this stage to support the resampling approach. Experience with computer simulations would suggest that much research would be needed in developing instructional activities or support for students to gain a deeper conceptual understanding of the randomization process, the notion of "chance" outcomes, and the distribution of the difference in medians to determine the likelihood of a particular outcome. At the present time the resampling approach to teaching would appear to be the most promising direction, as it could enable students to link probability intuitively with statistical inference. Basically, students need to understand that perceived patterns in data may be due to "chance", in which case inferential procedures should be conducted to determine whether the pattern is "real" or "random". Furthermore, a statement expressed in the language of probability, which is an assessment of the strength of the evidence for the correctness of the conclusion since the sample is not the entire population, must accompany any conclusion. There should be recognition that an element of uncertainty will always prevail.

5. CONCLUSION

How can teachers enable learners to make the connection between probability and statistical inference? Current research points to two strategies: first, emphasizing actual experience with exploring data before making connections between probability and inference; second, building a pedagogical framework, such as the one proposed in this chapter, to define a teaching pathway towards formal inference in Grades 9 to 12. Formal inference using a resampling approach could be introduced at Grade 12 although some educators suggest an introductory tertiary course is more appropriate. For the mathematically able students and those students whose careers will involve "t-tests", the resampling approach and the classical approach could possibly be integrated in tertiary level courses. Educators such as Cobb and Moore (1997) and Garfield and Ahlgren (1988) suggest that for a conceptual grasp of inference, informal probability is sufficient. Deriving the distributions and understanding inference from the classical viewpoint should be left for advanced study.

Historically full integration of probability and statistical inference in the statistics discipline only occurred in the first half of the twentieth century. Therefore it is not surprising that research on the teaching and learning of statistical inference with its inherent probabilistic nature is only in its infancy. Concerted efforts should be made by researchers to develop a teaching pathway towards formal inference as well as to investigate and develop new teaching approaches for formal inference. A key part of these investigations will be the linking of variation ideas and probability.

REFERENCES

Aspinwall, L., & Tarr, J. E. (2001). Middle school students' understanding of the role sample size plays in experimental probability. *Journal of Mathematical Behavior, 20*, 229-245.

Biehler, R. (1994). Probabilistic thinking, statistical reasoning and the search for causes: Do we need a probabilistic revolution after we have taught data analysis? In J. Garfield (Ed.), *Research Papers from The Fourth International Conference on Teaching Statistics, Marrakech 1994.* Minneapolis, MN: University of Minnesota.

Biehler, R. (1997). Students' difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 169-190). Voorburg, The Netherlands: International Statistical Institute.

Biehler, R. (2001, September). *"Girls (tend to) watch less television than boys" – Students' hypotheses and data exploration strategies in group comparison tasks.* Paper presented at LOGOS #10, Mathematics Education Unit, Department of Mathematics, The University of Auckland, New Zealand.

Biggs, J.B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy.* New York: Academic Press.

Cobb, G., & Moore, D. (1997). Mathematics, statistics and teaching. *The American Mathematical Monthly, 104*(9), 801-823.

Curcio, F. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education, 18*(5), 382-393.

Davis, P., & Hersh, R. (1986). *Descartes' dream.* Orlando, FL: Harcourt Brace Jovanovich.

DelMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education* *7*(3). [Online: www.amstat.org/publications/jse/v7n3]

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, *19*(1), 44-63.

Gravemeijer, K. (1998). Developmental research as a research method. In A. Sierpinska and J. Kilpatrick (Eds.), *Mathematics education as a research domain: A search for identity* (Book 2, pp. 277-295). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Joiner, B. (1994). *Fourth generation management.* New York: McGraw-Hill Inc.

Jones, P., Lipson, K., & Phillips, B. (1994). A role for computer intensive methods in introducing statistical inference. In L. Brunelli & G. Cicchitelli (Eds.), *Proceedings of the First Scientific Meeting of the International Association for Statistical Education* (pp. 199-211). Perugia, Italy: University of Perugia.

Key Curriculum Press Technologies (2000). *Fathom.* Emeryville, CA: Author.

Konold, C. (1994). Understanding probability and statistics through resampling. In L. Brunelli & G. Cicchitelli (Eds.), *Proceedings of the First Scientific Meeting of the International Association for Statistical Education* (pp. 255-263). Perugia, Italy: University of Perugia.

Konold, C., & Miller, C. (1995). *DataScope.* Santa Barbara, CA: Intellimation Library for the Macintosh.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259-289.

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 151-168). Voorburg, The Netherlands: International Statistical Institute.

Konold, C., Robinson, A., Khalil, K., Pollatsek, A. Well, A., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society*, Cape Town, South Africa. [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Lightner, J. (1991). A brief look at the history of probability and statistics. *Mathematics Teacher, 84*(8), 623-630.

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society,* Cape Town, South Africa. [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Lipson, K., Kokonis, S., & Francis, G. (2003). Investigation of statistics students' experiences with a web-based computer simulation. In *Proceedings of the International Association of Statistical Education Satellite Conference on*

Statistics Education: Statistics education and the internet, Berlin [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

McClain, K., Cobb, P., & Gravemeijer, K. (2000). Supporting students' ways of reasoning about data. In M. Burke & F. Curcio (Eds.). *Learning mathematics for a new century: 2000 Yearbook* (pp. 174-187). Reston, VA: National Council of Teachers of Mathematics.

Metz, K. (1999). Why sampling works or why it can't: ideas of young children engaged in research of their own design. In F. Hitt & M. Santos (Eds.), *Proceedings of the 21st annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 492-498). Cuernavaca, Mexico: PME.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: new approaches to numeracy* (pp. 95-137). Washington, D.C.: National Academy Press.

New Zealand Qualifications Authority (2001). *Level 1 achievement standards: Mathematic* [Online]. Available: http://www.nzqa.govt.nz/ncea/ach/mathematics/index.shtml

Pfannkuch, M., & Horring, J. (2004). *Developing statistical thinking in a secondary school: A collaborative curriculum development.* Unpublished manuscript, The University of Auckland, New Zealand.

Pfannkuch, M., & Wild, C. J. (2003). Statistical thinking: How can we develop it? In *Proceedings of the 54th International Statistical Institute Conference* [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Pfannkuch, M., Rubick, A., & Yoon, C. (2002). Statistical thinking: An exploration into students' variation-type thinking. *New England Mathematics Journal, 34*(2), 82-98.

Saldanha, L., & Thompson, P. (2001). Students' reasoning about sampling distributions and statistical inference. In J. Garfield, D. Ben-Zvi, & C. Reading (Eds.), *Background Readings of the Second International Research Forum on Statistical Reasoning, Thinking, and Literacy,* (pp. 291-296). Armidale, Australia: Centre for Cognition Research in Learning and Teaching.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257-270.

Scheaffer, R. (1992). Data, discernment and decisions: an empirical approach to introductory statistics. In F. & S. Gordon (Eds.), *Statistics for the twenty-first century.* (MAA Notes, Number 26, pp. 69-81). Washington, D.C.: The Mathematical Association of America.

Scheaffer, R., Watkins, A., & Landwehr, J. (1998). What every high school graduate should know about statistics. In S. Lajoie (Ed.), *Reflections on statistics: learning, teaching, and assessment in Grades K-12* (pp. 3-31). Mahwah, NJ: Lawrence Erlbaum Associates.

Shaughnessy, M. (2003). Research on students' understandings of probability. In J. Kilpatrick, W. G. Martin, & D. Schifter, (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 216-226). Reston, VA: National Council of Teachers of Mathematics.

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

Skovsmose, O., & Borba, M. (2000). *Research methodology and critical mathematics education* (Publication No. 17). Roskilde, Denmark: Centre for Research in Learning Mathematics, Roskilde University.

Stohl, H., & Tarr, J. E. (2002). Developing notions of inference using probability simulation tools. *Journal of Mathematical Behavior, 21*, 319-337.

Watson, J. M. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics, 47*, 337-372.

Watson, J. M. (2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics, 51*, 225-256.

Watson, J. M. (2004). Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 277-294). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Watson, J. M., & Moritz, J. B. (2000). Development of understanding of sampling for statistical literacy. *Journal of Mathematical Behavior, 19*, 109-136.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, 67(3), 223-265.

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

SECTION V

TEACHERS AND PROBABILITY