Comparison of data plots: building a pedagogical framework

Maxine Pfannkuch, Stephanie Budgett, Ross Parsonage, The University of Auckland, New Zealand
Julia Horring, Auckland Girls' Grammar School

This paper describes an emerging pedagogical framework for statistical inference for Year 11 (15 year-old) students in response to new assessment demands. As part of a three-year research project on developing statistical thinking in a school, half of the students identified, through an open-ended questionnaire, that they found it difficult to draw conclusions when comparing data plots. This paper focuses on this problematic situation by giving a brief analysis of student assessment task responses and learning opportunities in the classroom, hypothesising the reasons for the problem, and presenting a framework to help redress the situation.

Background

As part of the national mathematics curriculum (Ministry of Education, 1992) all students study statistics from Year 1 to Year 12, with many choosing to specialise in statistics at Year 13. Conducting statistical investigations using the whole empirical enquiry cycle (problem, plan, data collection, analysis, conclusion) are a core part of the curriculum for all levels. In 2002 a new approach to national assessment was introduced at Year 11. Instead of one final external examination in mathematics, one third of the course is internally assessed with the rest being an external examination (New Zealand Qualifications Authority, 2001). Since students are assessed on a full statistical investigation at Year 13 it was decided that Year 11 students would be given data sets to investigate with the emphasis on comparison of data and bivariate relationships. The level of statistical thinking required at Year 11 with this new internal assessment, compared to the previous external assessment which largely asked students to read and interpret graphs and calculate measures of central tendency, has placed real demands on teachers and students.

Previous research

Before students are introduced to confirmatory or formal inference methods to decide whether the patterns they see 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) analysed 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 described the language used by both teachers and students as inadequate. The problem of describing what is being communicated by a representation was also recognised by Bright and Friel (1998).

In Konold, Pollatsek, Well, and Gagnon's (1997, p. 165) analysis of the same Grade 12 students they hypothesised 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 was deemed problematic. McClain, Cobb, and Gravemeijer (2000), however, believed that their instructional experiments designed to develop seventh-grade students' notions of distribution, and argumentation which focussed on patterns in how the data were distributed, developed students' ability to reason about group propensities. This 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. Issues such as sample size and sampling variability in the argumentation 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 learnt from the data. Their idea of a data detective approach to data analysis fits with Pfannkuch, Rubick and Yoon (2002), who believe students should approach data analysis in the thinking roles of hypothesis generator, discoverer, and corroborator when working with data. Whilst not disagreeing with these recommendations the question remains as to when and how do you start building up concepts for formal inference.

Biehler (2001) argued that there was a four-stage development process for formal inference. In the context of an example involving the comparison of two boxplots he described the stages as: the EDA methods expert (fine tuning the comparison); the subject matter researcher and discoverer (widening and exploiting the context by bringing in more variables); the critical theory builder (generalisation); and the inferences statistics expert (Can group differences be "due to chance"?). These stages could be viewed as a learning pathway over time and as a four-stage approach to data analysis that would be expected from a senior student. The generalisation stage is fundamental to statistical inference in that there is 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. Recent research suggests that better teaching methods are needed to improve students' conceptual understanding of sampling in relation to statistical inference (Watson, in press). Research on the last stage of Biehler's model is limited. Efforts by Konold (1994) to improve students' understanding of the distribution of the mean differences for statistical inference using a resampling approach were somewhat effective. Lipson (2002) reported that students' understanding of sampling distribution with a particular software package was inadequate whereas delMas, Garfield, and Chance (1999) reported some success with their software. It would seem that the integration of statistical data analysis with theoretical probabilistic distributions and the assumptions underlying those models present a conundrum in teaching.

Research method

A developmental research method is used that is based on the ideas of Gravemeijer (1998), Wittmann (1998), and Skovsmose and Borba (2000) (Pfannkuch & Horring, 2004). The school selected draws on students from low socio-economic backgrounds, is culturally diverse, and has teachers interested in improving their statistics teaching. The teachers selected Year 11 (15 year-olds), and the case study teacher was self-selected. A workshop, which focussed on communicating the nature of statistical thinking (Wild & Pfannkuch, 1999) to the teachers, was conducted by the first author. After the workshop the case study teacher and another teacher were interviewed to identify problematic areas in their statistics-teaching unit (Pfannkuch & Wild, 2003). These two teachers and the first author then discussed teaching ideas that could be implemented to enhance the development of students' statistical thinking. The case study teacher then wrote a new four-week statistics-teaching unit. Although all Year 11 teachers implemented the new teaching unit research data were mainly collected from the case study classroom. Data collected 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. The first analysis of these data focused on the identified problematic area of informal inference, which led to a consultation group of five statisticians being formed to debate and discuss possible ways to progress.

The assessment task

The students were given a table of data showing the maximum temperatures of two cities Napier and Wellington, which were taken from some summer newspapers. 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?), analyse the data, draw a conclusion, justify the conclusion with three supporting statements and evaluate the statistical process. All students analysed the data by calculating the five summary statistics with many using back-to-back stem-and-leaf plots for these calculations and then drawing boxplots by hand. Figure 1 shows the boxplots drawn electronically. Note that Year 11 students are not expected to identify outliers so the whiskers were drawn to the minimum and maximum observations.

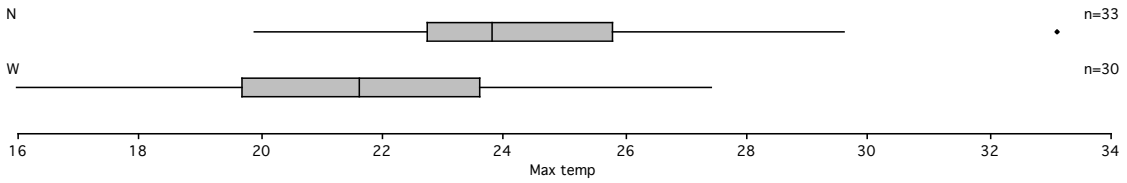

Figure 1. Comparison of Napier (N) and Wellington (W) maximum temperatures in the summer

Results

The analysis of the student assessment responses to the assessment tasks used a hierarchical performance level approach based on the SOLO taxonomy (Biggs & Collis, 1982). A spreadsheet was used to categorise the responses. Based on the student responses, four categories of justifications for their conclusions were identified: comparison of equivalent summary statistics; comparison of non-equivalent summary statistics; comparison of variability; and comparison of distributions. Within these categories hierarchies of responses were identified and qualitatively described. Generally the levels had the following characteristics: 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, the second author independently coded all responses. A consensus was reached between the first and second author on the final codes for each student response. The details of the student responses are recorded in Table 1.

Table 1: *Details of student responses when comparing boxplots*

	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

This analysis revealed that most students compared features in a non-discriminating manner, and did not justify or explain how their analysis supported their conclusion and was appropriate in relation to the question. For the boxplots, comparing similar summary statistics (27/30), including the range (16/30), which was not relevant to the question, were prevalent student responses. Eighteen students attempted comparison of non-equivalent summary statistics. There was no attempt at comparing the variability in relation to the difference in medians and little attempt at comparing the shape of the

distributions. Conclusions ranged from non-use of comparison language to comparisons that suggested statistical tendency. Figure 2 gives examples of student responses. 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.

Conclusion
• Napier has the highest temperature
• Napier is warmer than Wellington
• Napier tends to be warmer than Wellington.

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

Comparing non-equivalent summary statistics
• Napier's median temperature is higher than Wellington's upper quartile
• Also because the median of Napier's temperature is higher than three quarters of Wellington's temperature which suggest that half of Napier's temperatures are warmer than three quarters of Wellington's.

Comparing variability
• Napier has a larger range of data compared to Wellington
• 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.

Comparing distributions
• The box plot for Wellington is drawn lower than Napier's.

Figure 2. Examples of student responses

When the students evaluated the statistical process 20/30 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.

Discussion

Hypotheses were generated as to why drawing a conclusion and justifying it were problematic when comparing data plots. One hypothesis was that texts and therefore teaching tended to compare only features of boxplots and to *not* draw a conclusion, since significance testing and confidence intervals are introduced at a later stage (Wild & Seber, 2000). Other hypotheses were that the assessment demands were beyond the capabilities of Year 11 students, that 'informal inference' techniques are not established or recognized within the statistics discipline, that the curriculum does 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, it was the inherent messiness of data, the absence of a clear decision, and the positing of possible contextual explanations, that made data comparison interesting. If informal

inference was to be taught there might need to be more awareness amongst teachers of the formal inference ideas underpinning comparison of data plots.

From the perspective of formal inference for the comparison of data plots the statisticians determined that there were four basic aspects to attend to in order to understand the concepts behind significance tests, confidence intervals, p-values and so forth before drawing a conclusion. These were: comparisons of centres; taking the variability into account relative to the differences in the centres; 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 for comparison of data plots for Year 11 is beginning to be developed. This framework, based on the assumption that formal inference notions should begin to be developed by Year 11, will continue to be under debate amongst the teachers, researcher, and statisticians. It is a framework for making teachers aware of the 'big ideas' that students need to experience and develop for inference, namely, (i) knowing why they should compare centres, (ii) describing and interpreting variability within and between sample distributions, (iii) developing their sampling reasoning, and (iv) how to draw an acceptable conclusion based on informal inference.

Comparing the centres

Wild and Pfannkuch (1999, p. 240) said that "the biggest contribution of statistics is the isolation and modelling 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. One third of the students recognized that the comparison of the medians only was one justification for their conclusion. According to Konold and Pollatsek (2002, p. 273) statistical reasoning will elude students unless they understand that the comparison of averages is the statistical method for determining whether there is a difference between two sets of data and that "this pattern is symptomatic of students' failure to interpret an average of a data set as saying something about the entire distribution of values". It would not be obvious to these students why the comparison of centres 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 shape. These ideas based on the Galton board should be extended to include drawing two graphs of the same data. First, students should construct a series graph to visualise 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 non-intuitive way.

Variability, checking and comparing the sample distributions

When comparing variability half of the students compared the ranges, which was not relevant to the question. Formal inference requires comparison of the variability relative to the difference in the centres, which presupposes an understanding of standard deviation or confidence intervals. A statistician might 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 Year 11 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 should be on describing, interpreting, and comparing the variability in the data sets not on using it to answer a question that asks for "the warmer" or "the better" and so forth. 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 is a noteworthy feature to describe.

Sampling reasoning

A major problem with informal inference is taking sample size into account. There are many strands to building up concepts about sample size effect. From this research some matters that need to be attended to are: comparison of sample data of unequal sample size; the notion of a sample; small sample versus large sample variability; the sample and its relationship to the population; and the size of the sample and its relationship to the population. A repertoire of teaching and learning possibilities needs to be considered to build up these concepts (Watson, in press).

For these students it is 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 first author's observation, corroborated by the teachers, was that the students had experience of reading the data, not much 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 learnt that their predictions were likely to be within the interquartile range or at least within the range. Missing data is a well-known problem in statistics and students should be given opportunities to impute values for observations and to analyse data with and without the imputations. Specific attention should be drawn to students' beliefs and to whether their conclusions would change with unequal sample sizes.

This problem is compounded by sample size and variability being interconnected. Simulations such as taking random samples of the same and different size from a population to 'see' the variability of the sample mean, and the variability within and between samples should be part of students' learning. Meaningless simulations would not advance students' conceptions of the sample size effect. Hypothetical situations grounded within the context of a problem (e.g. If they took another summer's temperatures would they get the same graphs?) might start to induct students into some formal inference ideas.

Drawing a conclusion

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 when plots are considered to be overlapping. Simulations could be used to overcome students' beliefs that "50% of Napier's temperatures are higher than 75% of Wellington's temperatures" is evidence for a real difference and that the sample size of 30 is large enough. For example, students' attention could be drawn to noticing that some of the randomly generated plots of sample size 30 from the same distribution would give rise to the above statement. The simulations should generate boxplots and histograms, as these are the types of graphs from which the students are required to make informal inferences.

To remedy the Year 11 assessment requirements of drawing a justified conclusion it was suggested that students "look at the plots" and compare the centres, spreads, and anything else that is noteworthy. After comparing and describing features, students should then draw an informal inference along the lines: "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 formal inference notions. The conclusion could be justified by referring to the comparison of centres. The question of whether the students are drawing a valid conclusion is being addressed in another framework that focuses on the evaluation of the statistical process.

The building of a pedagogical framework, in response to a problematic situation, should be viewed as a sub-framework that interconnects with other statistical frameworks at Year 11 and across all Year levels. In relation to Biehler's (2001) four-stage model the Year 11 experience should be building up concepts for the third stage, the critical theory builder. The framework will 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 inferential reasoning will continue to elude most students.

References

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

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

Bright, G., & Friel, S. (1998). Graphical representations: Helping students interpret data. In S. Lajoie (Ed.), *Reflections on statistics: learning, teaching, and assessment in grades K-12* (pp. 63-88). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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), www.amstat.org/publications/jse/v7n3

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* (pp. 277-295). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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. 199-211). Perugia, Italy: University of Perugia.

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.

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.

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* (pp. 174-187). Reston, VA: National Council of Teachers of Mathematics.

Ministry of Education (1992). *Mathematics in the New Zealand Curriculum*. Wellington, New Zealand: Learning Media

New Zealand Qualifications Authority (2001). *Level 1 achievement standards: Mathematics*. [Online]: 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 paper.

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

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.

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

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

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

Wild, C. & 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.

Wittmann, E. (1998). Mathematics education as a 'design science'. In A. Sierpinska and J. Kilpatrick (Eds.), *Mathematics education as a research domain: A search for identity* (pp. 87-103). Dordrecht, The Netherlands: Kluwer Academic Publishers.