STUDENTS' EMERGING EXPRESSIONS OF UNCERTAINTY WHILE MAKING INFORMAL STATISTICAL INFERENCES ABOUT DATA

PRODROMOU, Theodosia
University of New England,
Australia

Contact email: theodosia.prodromou@une.edu.au

ABSTRACT

This research study investigates the development of middle school students' emerging expressions of uncertainty through observation of 14- to 15-year-olds, challenged in informal inferential reasoning. This study focuses on students' investigations when sampling from populations and using information from the samples to draw conclusions about the parent populations. The results suggest that when the students engaged in processes of drawing generalised conclusions from data, involving generalising beyond data and using data as evidence of the generalisation, they developed probabilistic language to articulate the degree of certainty embedded in the generalisation. As the students engaged in their inquiries, they developed more sophisticated expressions of the probabilistic language. Attending to students' emerging articulations of uncertainty when making judgments about the underlying structure of the data and observing patterns and trends in data, provides an opportunity to develop more sophisticated understandings of the developmental process of students' statistical inferential reasoning.

INTRODUCTION

Statistical inference is the process of drawing conclusions about populations, using datasets drawn from the population of interest via some form of random sampling that is subject to random variation, such as observational errors, or sampling variation. In essence, statistical inference aims to infer an unknown parameter for a given population, based on a sample taken from the population. The aspects studied by statistical inference are divided into estimation and hypothesis testing. We employ the term 'statistical inference' to refer to a group of common forms of statistical schemes addressed by an estimate (i.e., a particular value that best approximates a parameter of interest); a confidence interval (i.e., an interval constructed using a dataset drawn from a population so that, under repeated sampling of such datasets, such intervals would contain the true parameter value with the probability at the stated confidence level); critical values; p-values; and posterior distributions. More generally, statistical inference deals with the specific type of uncertainty caused by having only data from random samples obtained from populations rather than having data from the entire population, process, or distribution.

This paper focuses on students' reasoning about, and conceptions of, statistical inference. Initially, we build on what is already known about students' intuitive reasoning. Such intuitive reasoning about statistical inferences has been described as informal inferential reasoning, a process of making generalisations, including conclusions, predictions and estimating parameters beyond describing the given data (Makar & Rubin, 2009), and comparing datasets to a conceptual model (Bakker et al., 2008). Moreover, the following principles were identified by Makar and Rubin (2009) as essential for informal statistical inference: the use of data as evidence for making generalisations and the employment of probabilistic language in describing the generalisation, including informal references to levels of certainty about the conclusions drawn. Research on informal statistical inference has been conducted on tertiary settings (Zeiffler et al., 2008), secondary (e.g., Prodromou, 2013a, 2013b; Watson, 2008), and primary schools (Makar & McPhee, 2009).

Inspired by Wild et al. (2011), who identified "a minimal set of the biggest ideas of statistical inference" and "integrated inferences for beginners within a holistic view of the investigative cycle (Wild and Pfankuch, 1999)," this paper adopts a "wholly visual approach" (Wild et al., 2011, p. 252) that will, attempt to minimize the conceptual distances between the concrete realities, including precursor practical experiences, and the dynamic imagery envisaged. Our ideal, moreover, has the

inferential step able to be performed without students taking their eyes of their graphs so that the connections between question, data and answers are kept as immediate and possible as possible. (p. 252)

It is hoped that this approach will help provide more intuitive connections to the more formal methods to be built and used later.

Students cannot make statistical inferences without an appreciation of the difference between description and inference. Whereas descriptive statistics are used to describe a sample's characteristics, inferential statistics are used to infer something about the population based on the sample characteristics.

These characteristics together with sampling variation are the critical elements that statistical inference was developed to deal with. In essence, a hallmark of any conceptual approach to statistical inference must derive from a good understating about the nature and behaviour of sample variation. The term 'sample variation' denotes the variation derived from the selection of a sample that represents their parent population to a greater or lesser extent. When making inferences, one attempts to account for the uncertainty due to having only a sample rather than having data from the whole population. One can obtain an idea of the extent to which sample data are likely to represent the parent population by looking at the properties of the data (e.g., by computing means or other statistics) over repeated samples or by examining the patterns of sampling variation. The uncertainty caused by sampling variation and the degree of uncertainty allowed for are estimated when studying patterns of sampling variation. Understanding such sampling variation and the degree of its uncertainty are of major importance for making decisions about populations in all empirical sciences, including psychology, healthcare, and education (Belia, et al., 2005).

THEORETICAL FRAMEWORK

Statistical inference, including use of hypothesis tests and confidence intervals, is one of the primary aspects of most curricula in statistics. However, students are usually prone to fall into many misconceptions when making statistical inferences (Kirk, 2001) because inferential statistics involve understanding many abstract concepts such as sampling distributions and significance levels. Research has shown that many students often are unable to integrate the fundamental concepts in inferential reasoning (Batanero, 2005). In particular, Chance et al. (2004) suggested that large numbers of students had difficulties comprehending sampling variation of means when using computer animations.

This has led researchers to suggest that,

> since statistical inference (used in its conventional sense) is designed to deal with uncertainties about the true state of nature due to sampling variation, we believe that experiences that are designed to build and cement the ideas of statistical inference should focus solely on sampling variation. (Wild et al., 2011, p. 253)

Wild and his team (2011) pointed out that variation such as 'random measurement error' mainly focuses on planning and critiquing investigations and not on introducing the core ideas of statistical inference. Research on informal statistical inference has reported that teaching experiences that require users to interpret the "centre differences" have often employed on context matter knowledge (Watson, 2008) throwing yet another complication into students' consideration of description versus inference. Wild et al. (2011) argued that critiquing the plausibility of an inference draws on knowledge of context, as does any attention to the practical importance of any differences seen among the patterns in data. To avoid any complications introduced by integrating knowledge of context and knowledge about inference, they propose that pedagogical exercises on informal statistical inference be visually oriented, drawing solely on patterns in data that are not closely linked to context, decision logic, and probability. Wild et al. (2011) propose decision guidelines that were devised to support the New Zealand high school statistics curriculum. Their decision guidelines involve 4 milestones, with students reaching one milestone per year of schooling, and with milestone 4 being targeted for the last year. The proposals of Wild et al. (2011) focused both on conceptual flow and on classroom implementations presented in terms of sampling from populations. Their proposals are built from particular conceptions of sampling variation built by using animated, computer-simulation-based boxplots.

They consider that boxplots provide a natural "bridge between reasoning entirely from graphics to reasoning from summaries in ways that converge, qualitatively, to the two-sample t-test" (p. 254) between operating entirely in terms of what is seen in graphics to reasoning using summaries The latter, are graphically depicted in a basic boxplot (box-and-whisker diagram or plot) creating the visualisation of the shape of the distribution, its central value, and variability.

Wild et al. (2011) tried to convey to their readers the sampling variation by using animated computer-based boxplots and figures intended to convey ideas about how to read inferential information from the box blots. The visualisation of sampling fluctuation is compiled by the superposition of repeated boxplots. The authors reported on 4 milestones.

Milestone 1 involves (a) an appreciation that samples can provide us with useful approximate pictures of populations, (b) an ability to observe approximate location changes in boxplots, (c) an appreciation that the story told by the data about the population can be wrong, and (d) an appreciation that the shift that is observed in data must be reasonably significant before we can fairly safely infer the direction of a population effect from the direction of a data effect.

At milestone 2, all of the milestone 1 points (a)-(d) should be reinforced. We are no longer just concerned with the relationship between a single sample and a single population, but that now we're talking about multiple samples from the same population, or even multiple samples from multiple populations. Milestone 2 adds two new requirements to the first milestone: first, the sample size should be taken into account when seeing shifts of canters in data, and, second, a change of attention towards distance between centres as a proportion of a spread. The authors first attempted to compare the distance between medians with the sum of the interquartile ranges but they were informed by the teachers that it was too difficult for their students, thus this conversation led researchers to the 'overall visible spread' idea. They obtained the cut-off proportions that are depicted by using simulations with normal data.

The type I error rates are about 8% at the anchor sample sizes. There was a trade-off between more conventional type I error rates at memorable sample sizes and having a simple rule. For example, the round number sample sizes with approximate 5% type I error rates are n=40 for $\frac{1}{3}$, n=80 for $\frac{1}{4}$ and n=125 for $\frac{1}{5.}$. The type I error rates with data from the strongly skewed or heavy tailed distribution are similar at the anchor sample sizes to those from the normal distribution.

Milestone 3 continues the convergence towards the big idea of the t-statistic. They used 1.5 multiplier, increasing the large sample type I error rate with normal data slightly from about 2% to about 2.5%. Additionally they drew a thick horizontal line in place of a notch. This is approximate 90% confidence intervals. They then used the non-overlap of individual uncertainty intervals to show significance. They believed that by operating this way, they worked with Type I error rates for significance tests for a difference that ate much smaller than the convergence error rates for the individual parameters.

Milestone 4, brings in the notion of null hypothesis, levels of variational behaviour under the null hypothesis due to sampling or randomization, normal distribution models, alteration of emphasis for measures of location and spread from the median and interquartile range to the mean and standard deviation, and formal methods of inference based on t-tests and randomization.

Borovcnik (2011) mentioned the poorer behaviour of the milestone 3 rule that is caused by the 'intuitive intervals' that treat extreme and usual cases as equally likely. He found milestone 3 to not be a smooth progression to a more conventional milestone 4 of formal statistical inference because the procedure widens the confidence interval precluding a probabilistic argument. This paper, instead of dismissing that milestone 3 is not a smooth progression to a more conventional milestone 4 of formal statistical inference or eradicating students' difficulties in constructing probability arguments; it considers students' difficulties as starting points that provide a pedagogical challenge of how to build on learners' impoverished understandings when making statistical inferences. This research study is interested in investigating how students reason about the intervals for the median and make probabilistic arguments about the difference that there is back in the populations.

METHODOLOGY

Thirty students in Year 9, ranging from 14 to 15 years in age, from a rural secondary school in New South Wales, Australia, formed the population of this study. The researcher spent 2 sessions (40-45 minutes each) introducing the class teacher and the students to Geogebra4.2 during regular mathematics lessons. All students were familiarised with the Geogebra4.2 software, explicitly focusing on learning skills related to the software.

In the first session, all students were familiarized with Geogebra4.2 software through a number of introductory activities related to creating statistics summaries, graphing data, and comparing data sets. In the second session, all students analysed data about students' (Years 7−9) weight of backpacks. The students of the research study were asked to compare the students' (Years 7−9) weight of backpacks and how it differs compared to the national average weight (5 kgs.) of students' (Years 7−9) backpacks. In this study, we observed students to compare female and male students' (Years 8) weight of backpacks. The student participants were asked to limit themselves to sample sizes of around 20-40 students (one classroom for each year). This requirement had the advantage of simplifying the procedure at the cost of limiting its utility.

The participants used features of Geogebra4.2 to construct two parallel Boxplots (Figure 1). The two boxplots summarize data from Year 8 students' weight of back bags. Whereas Boxplot A summarizes data from girls' (Year 8) weight of backpacks, Boxplot B summarizes data from boys' (Year 8) weight of backpacks.

The qualitative research study was questioning-based and observation-based. The researcher was a participant observer who asked students questions and probed into students' reasons or intuitions that might explain their actions. The participant students were asked to reason about the intervals for the median and make a probabilistic argument about the observed difference that

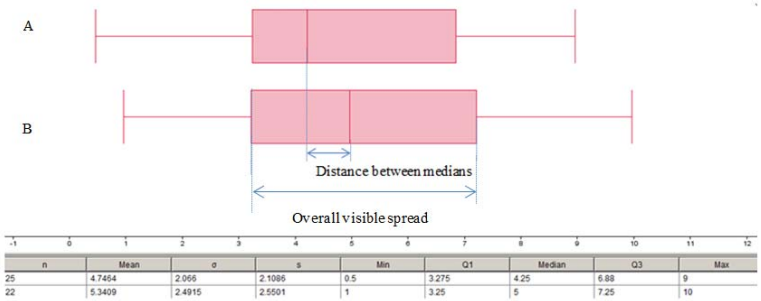

Figure 1: Boxplot A summarizes data from girls' (Year 8) weight of backpacks, Boxplot B summarizes data from boys' (Year 8) weight of backpacks.

there is back in the populations. While the students were working, Camtasia software was used to video record of the computer screen output and audio record the students' voices. At the first stage, the audio recordings were simply transcribed and screenshots were incorporated as necessary to make sense of the transcription.

Subsequently, the data were then analysed using progressive focusing (Robson, 1993), a process by which the author began with a wide field of focus and gradually narrowed the field by identifying key foci for ensuing study. A challenge that arose in such a qualitative research study was that of coordinating multiple levels of analysis. The effects of these threats were attenuated only by careful attention to the multi-faceted concepts of validity (e.g. content validity, construct validity, internal validity, external validity, concurrent validity, ecological validity, interpretative validity, theoretical validity etc. (Cohen et al., 2000, p. 106)) and reliability throughout this research study. In addition to generalisability as a criterion to external validity, the researcher focussed on certain common threads of students' articulations and looked for the applicability of those threads across the 30 students' activity. Those common issues may appear to readers as transferable to other settings, and situations.

RESULTS AND IMPLICATIONS

A qualitative analysis by the researcher of her discussion with the students on the comparison of box plot distributions was performed. Students observed that the median for the sample of girls' (Year 8) weight of backpacks was 4.25 kgs and the median for the sample of boys' (Year 8) weight of backpacks was 5 kgs. To operate the milestone 2 test, students did a quick freehand subdivision of a line representing total visible spread into thirds and make the decision on that basis. They found the $\frac{1}{3}$ of overall visible spread to be 1.71 and concluded that A tends to be bigger than B back in the populations because the distance between the medians that is 0.75 is smaller than about the $\frac{1}{3}$ of overall visible spread.

The students followed the Level 7 (Milestone 3) guideline to construct informal confidence intervals that capture the population medians (Figure 2). They calculated the interval estimates (confidence interval) for the population medians for A and B. They found the confidence interval for the population median for A to be (3.169, 5.333) and for B to be (3.721, 6.279).

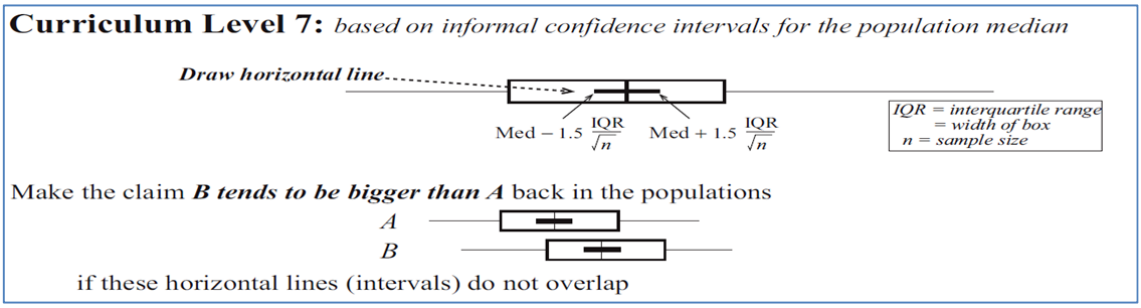

Figure 2: Development of formula for confidence interval (Milestone 3)

The researcher asked students explain their thinking about the confidence intervals for the population medians. The students articulated that "the confidence intervals contain all the medians and the actual median is likely to be in the confidence interval". The researcher continued with asking students what they meant by "is likely to be". The students replied: "I am not very sure… uncertainty is involved". The researcher asked students to explain how uncertainty is involved. The students replied: "Several factors affect the uncertainty of estimating the confidence interval for the population median such as sample size, the amount of variability in the population being study or spread of population, estimated with sample IQR". Another student added that he wanted to be confident that the interval estimate contains the true population median. The researcher asked: "How confident do we want to be that our interval estimate contains the true population median?" The student replied: "we have already mentioned three factors that influence confidence intervals: Sample size, variability in the populations. I will never be 100% confident. I will be confident if I know that we predict the population median 90% of the time." The researcher asked the students to explain. The students very well-articulated: "The interval includes the true population median for 9 out of 10 samples - the population median is probably in the interval somewhere."

The researcher shifted students' attention to sampling variation, asking students whether the sampling variation can produce a shift large enough so that the students can make a mistaken claim. Students explained that,

> when the populations do not overlap, they are able to make a claim about the populations. When the calculated intervals do not overlap, a confidence interval for the difference in the population medians ranges from the smaller distance between the intervals to the larger distance between the intervals.

Another student added that they were looking for sufficient evidence, a big enough shift in the intervals for the median to be able to make a claim that there was a difference back in the populations. The researcher asked the student what he meant by "sufficient evidence". The student explained:

> Sufficient evidence includes all possible significant results. This means that our data provides us with insights that give 'something out of the ordinary'… something that

> would have had a very small probability of happening just by chance… Or when you compare data sets, the difference in the groups of data is so big that it would be hard to say it was just a coincidence.

The researcher asked students to explain their articulation by giving an example. The students articulated:

> When we compare two drugs that treat cancer...If a drug is found to be more effective at treating cancer that the current treatment is, we can tell that the new drug shows improvement in the survival rate of patients with cancer. That means that based on data, the difference in the results from patients on the new drug compared to those using the old treatment is so big that it would have hard to say it was just a coincidence.

The results of implementing milestones 1-3 showed that students were able to reason about possible features of a population based on a sample of data drawn from the given population. The collected data also indicates that participants can use their informal statistical knowledge to reason about possible differences between two populations based on observed differences between two samples of data.

The present study provides new insights in participants' intuitive reasoning about the differences due to an effect as opposed to differences due to chance. In this study, students appreciated whether a specific sample of data is likely to have a particular characteristic that is being studied under a particular claim. For example, students reasoned about whether the actual population median is likely to be in the confidence interval.

The students of this research study found the procedure of calculating the confidence intervals counterintuitive statistically and somehow misleading because the extreme and usual cases were treated as equally likely. The results suggested that students articulated colloquial notions of chance reasoning that expressed uncertainty of conclusions drawn from sample data in the light of variability.

On limitation of this study is that no sessions followed up to assess whether the students could apply these approaches to new situations at a later time. Future research needs to investigate more systematically students' informal inferential reasoning about the uncertainty of claims made from sample data. Further search needs to clarify what aspects of formal inference are needed given current tools, approaches and methods to foster students' ability to use and understand conceptions of the more conventional milestone 4 of formal statistical inference.

REFERENCES

Bakker, A., Kent, P., Derry, J., Noss, R., & Hoyles, C. (2008). Statistical inference at work: Statistical process control as an example. *Statistics Education Research Journal*, *7*(2), 131-146.

Batanero, C. (2005). Statistics education as a field for research and practice. In *Proceedings of the 10th international commission for mathematical instruction*. Copenhagen, Denmark: International Commission for Mathematical Instruction.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand intervals and standard errors bars. *Psychological Methods, 10, 389-396.*

Chance, B., delMas, R. & Garfield, J. (2004). Reasoning about sampling distributions. *In the Challenge of Developing Statistical Literacy, Reasoning and Thinking* (Eds. D. Ben-Zvi and J. Garfield), pp. 295-324. Dordrecht: Kluwer.

Cohen, J., Manion, L., & Morrison, K. (2000). *Research Methods in Education*. London and New York: Routledge Falmer.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement,* 61 (2), 213-218.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, 8(1), 82-105.

Makar, K., & McPhee, D. (2009). Young children's explorations of average in a classroom of inquiry. In R. Hunter, B. Bicknell, & T. Burgess (Eds.), Crossing divides: *Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia*, p. 347-354. Palmerston North, NZ: MERGA.

Prodromou, T. (2013a). Estimating Parameters from Samples: Shuttling between Spheres. *International Journal of Statistics and Probability*, 2(1), 113-124. doi: 10.5539/ijsp.v2n1p113

Prodromou, T. (2013b). Informal Inferential Reasoning: Interval Estimates of Parameters. *International Journal of Statistics and Probability*, 2(2), 136-152. doi: doi:10.5539/ijsp.v2n2p136

Robson, C. (1993). *Real World Research*. Oxford: Blackwell.

Watson, J. M. (2008). Exploring beginning inference with novice grade 7 students. *Statistics Education Research Journal*, 7, (2) pp. 59-82.

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

Wild, C. J., Pfannkuch, M., Regan, M.,& Horton, N. J. (2011). Conceptions of Statistical Inference. In J. R. Statist. Soc. A, 174, Part 2, 247-295.

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, 7(2), 40-58.