DESCRIBING DISTRIBUTIONS

Pip Arnold[1,2] and Maxine Pfannkuch[2]
[1]Cognition Education Limited, New Zealand
[2]The University of Auckland, New Zealand
parnold@cognition.co.nz

Describing distributions require students to simultaneously consider many aspects of five overarching statistical concepts –contextual knowledge, distributional, signal and noise, variability and graph comprehension– making describing distributions challenging. In this paper two students descriptions of distributions are compared and contrasted through the lens of the distribution framework (Arnold & Pfannkuch, 2012) as an initial insight into the different features of distribution that are emphasised in teaching and learning within a year 10 class (ages 14-15).

INTRODUCTION

"'Distribution' is another fundamental given of statistical reasoning. I can find a great deal written about specialised usages and definitions of 'distribution' but almost nothing about 'distribution' itself as an underlying conceptual structure" (Wild, 2006, p.10).

Distribution is a multi-faceted notion involving centre, spread, skewness, shape and density (Bakker, 2004a; Ben-Zvi & Amir, 2005; Konold, Higgins, Russell, & Khalil, 2004; Pfannkuch, 2006; Reading & Reid, 2006). As they consider distribution, students will consider the following together to get the "bigger picture": measures of centre (mean and median); measures of spread (range, interquartile range and standard deviation); where the majority of data values are in relation to extreme values (skewness); and how density and skewness provide detail about shape. Knowledge about the context is another key aspect of the notion of distribution. This simultaneous consideration of many aspects requires global reasoning, the coordination of which makes distribution a complex notion that students across the system find difficult (Ben-Zvi & Arcavi, 2001; delMas, Garfield, & Ooms, 2005; Hancock, Kaput & Goldsmith, 1992; McClain & Cobb, 2001; Pfannkuch, 2006).

Distribution is most often realised in a display such as a graph. Friel, Curcio and Bright (2001) recognise the complexity of reading graphs and what the graph reader must attend to; for example, describing, representing, analysing and interpreting data as well as taking the context into account. Questions such as "How does the learner's understanding of the context contribute to his or her interpretation of data represented in a graph?" and "Can one interpret data accurately without having a significant level of understanding of the context?" (Friel et al., 2001, p. 152) are challenges to think deeply about when considering the role context plays in students' understanding of distribution. Bakker (2004a) and Watson (2005) also note the importance of context both in recognising shape and in understanding how much variation from an expected distribution is realistic. A related matter is the inclusion of the context in the description of the distribution. Without the contextual references, the description becomes meaningless. The complete description of a distribution is "one that involves, shape, center, and spread, and in the context of the data" (delMas et al., 2005, p. 6). These features in the description are properties of the aggregate or collection as a whole and not of individual data values (Konold et al., 2004).

An end goal therefore is that students are able to describe sample distributions as part of their using the statistical enquiry cycle to answer an investigative question about a population. The following quotes encapsulate big ideas around distributions and why being able to understand the component parts and describe them in context is important. "The power of statistical data analysis lies in describing and predicting aggregate features of data sets that cannot be noted from individual cases" (Bakker, 2004a, p. 100) and "A larger objective in terms of the goal of statistical literacy when students leave school is to be able to tell a story from a context with a graph that displays variation, clustering, middles and surprises" (Watson, 2005, p. 189).

While some researchers, such as Bakker (2004a, 2004b), have reported on students' responses regarding describing shape, no one had explored how to develop students' language for describing distributions and what describing data distributions encapsulates, in particular for 14- and 15-year old students.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

DISTRIBUTION FRAMEWORK

As part of the first author's doctoral thesis describing distributions was explored in response to answering statistical investigative questions (Arnold, 2013). Findings from this exploration included the distribution shape descriptors students (year 10, ages 14–15) intuitively use to describe the shape of a distribution (initial reporting in Arnold & Pfannkuch, 2012), what makes a good distribution description at ages 13–15 (New Zealand curriculum level 5, also in Arnold & Pfannkuch, 2012) and the distributional shapes and graphs students (year 10, ages 14–15) predict when given the context.

From this analysis of student work and the literature a distribution framework (Figure 1) for curriculum level 5 (ages 13–15) was developed (Arnold, 2013; Arnold & Pfannkuch, 2012) to support describing distributions. The framework, however, includes all aspects of distribution not just descriptions. The Bakker and Gravemeijer (2004) proposed structure for analysing the relationship between data and distribution, the Ben-Zvi, Gil and Apel (2007) theoretical framework for informal inferential reasoning, and the Pfannkuch, Regan, Wild and Horton (2010) ideal data-dialogue were all used to build the basis of the distribution framework. Additional characteristics and features (noted in italics in Figure 1) were the result of the analysis of student work.

Overarching statistical concepts	**Characteristics of distribution**	**Specific features measures/depictions/descriptors**
Contextual knowledge	Population	1. *Target population* (e.g. New Zealand year 5–10 students) 1. Other acceptable population (e.g. year 5–10 students)
	Variable	2. Variable 3. Units 4. *Values*
	Interpretation	5. *Statistical feature described in contextual setting* (e.g. interpreting right skew as very few high test scores, with most test scores between 20 and 50 points)
	Explanation	6. *Possible reason for a feature* (e.g. bimodal due to gender for kiwi data)
Distributional	Aggregate view	7. General shape sketched (correctly) 8. Hypothesis and prediction
	Symmetry	9. Overall shape
	Modality	10. *Modality*
	Skewness	11. Position of majority of the data (to the left or the right)
	Individual cases	12. *Highest and lowest values*
Graph Comprehension	Decoding visual shape	9. Overall shape 13. *Parts of the whole* (splitting the distribution into parts and describing the parts as well as the whole) 10. *Modality*
	Unusual features	14. Gaps 15. Outliers
Variability	Spread	16. Range, 17. interquartile range 18. *Range as an interval* (e.g. heights range from 132-197 cm) 19. *Interval for high and/or low values (*may be describing a tail) 20. *Interval for groups* (e.g. in the case of a bimodal distribution)
	Density	21. Clustering density 22. Majority (mostly, many) 23. Relative frequency
Signal and noise	Centre	24. Median, 25. Mean
	Modal clumps	26. *Peak(s)* 27. *(local mode)* 28. *Modal group(s)*

Figure 1. Distribution framework for curriculum level 5 (ages 13–15)

Having identified the 28 different distribution features that are possible at this level (numbered in Figure 1), a secondary analysis on the work of two year 10 (ages 14–15) students was conducted. Hence the research questions for this paper are (1) to what extent are two year 10 students using the different distribution features in their pre- and post-test and homework responses? and (2) how are these two year 10 students using the different distribution features in their pre- and post-test and homework responses?

METHOD

The research method followed design research principles (Roth, 2005) for a teaching experiment in a classroom. In the preparation and design stage the first author developed the teaching and learning materials for a 16-lesson teaching experiment in conjunction with the classroom teacher, considering relevant literature. Both the classroom teacher and first author were involved in the implementation of the activities in the teacher's year 10 (ages 14–15) class. Following each lesson there was reflective discussion and adjustments were made as needed to the learning trajectory. The learning activities were designed to support students' understanding of describing distributions and built on research work previously undertaken (Pfannkuch, Arnold, & Wild, 2011). New activities were developed and were based on the themes that emerged from the literature. In particular the activities focused on the language of shape, making predictions, building a contextual knowledge base for the sorts of variables that have symmetric, skewed or uniform distributions, all of which support students' descriptions of distributions.

Describing distributions were the specific focus of lessons 2–4. The learning trajectory was designed to specifically support students developing an idea of what is meant by: (1) shape in relationship to distributions, (2) sketching and describing the shape of a distribution, (3) predicting the shape of distributions, (4) identifying and describing features of data distributions, and (5) connecting all of these to the context, i.e. to both the population(s) and the variable(s). In lesson 4 the focus was on describing the distribution of a graph and from that lesson onwards the students were given a different graph each night for homework for which they were required to write a description. One description was shared at the beginning of each lesson and the teacher actively reflected on the student's description to improve and complete the description as necessary.

The 29 students in the class were above average in ability and from a mid-size (1300), multicultural, mid socio-economic inner city girls' secondary school. Students were given a pre- and post-test, the lessons were videotaped and student work was photocopied. A group of six girls were observed specifically as well as the teacher led whole class discussions. The six girls also had pre and post-interviews about their responses to their tests. The girls were chosen on the basis that they gave consent to be videoed and interviewed and that they already sat as a group in the class.

The two students (student A and student B), whose responses are analyzed for this paper, were selected from this group of six. They were chosen as they represented the two extremes in terms of improvement in performance from pre- to post-test. Their pre-and post-test distribution descriptions were graded using the SOLO taxonomy (Biggs & Collis, 1992; Watson, 2005). The particular descriptors for describing distributions were developed through a process of moving between the literature, in-class observations and student responses. In brief the descriptors for grading the student responses were: *no response* (NR-0); *pre-structural* (PS-1) – context and/or evidence missing; *uni-structural* (US-2) – gives one correct piece of evidence in simple context OR multi-structural evidence without any context; *multi-structural* (MS-3) – identifies a simple context and correctly describes two features OR relational evidence without any context; *relational* (R-4) – identifies the context, connects the context, and correctly describes the overall shape and at least two other features; *extended abstract* (EA-5) – identifies the context, connects the context throughout the description, correctly describes the overall shape and at least three other features and may include some explanation or interpretation of results to the context. Based on this grading system student A's grade from pre- to post-test remained the same (MS-3) while student B's grade moved from uni-structural (US-2) to extended abstract (EA-5).

RETROSPECTIVE ANALYSIS

The two students' pre- and post-test and homework responses were analysed for this paper. The focus was on the actual features that the two students chose to use in their descriptions for each of the three graphs in the pre- and post-test and for the 12 graphs for homework. Altogether 18 descriptions were analysed for each of the two students.

In the pre- and post-tests students were asked to write two statements about the distribution of the variable for each of the three situations given. Generally they wrote more than the required two statements. In the pre-test student A used seven different features (1, 2, 3, 4, 11, 18, 25) and student B used six different features (2, 3, 4, 11, 18, 27) across the three descriptions. Collectively the two students used eight of the 28 features (see Figure 1). In the post-test student A used eight different features (4, 6, 7, 9, 10, 13, 21, 26) and student B used 12 different features (1, 2, 3, 4, 7, 9, 10, 13, 18, 21, 24, 26) across the three descriptions. Collectively the two students used 13 of the 28 features in the post-test.

The two students had used the mean (25) and the mode (27) in the pre-test, but not in the post-test, reflecting the use of the median (24) to describe the centre in the class teaching. They had also noted the position of the majority of the data (11) in the pre-test and not in the post-test. Additional features mentioned in the post-test included the overall shape (sketched (7) and described (9)), modality of the shape (10), describing parts of the whole (13), clustering density (21), median (24), peaks (26) and a possible reason for the feature (6). All of the additional features used in the post-test that were not used in the pre-test reflect work that they did in class.

The post-test descriptions for student A were of a much lower quality than her homework descriptions (see examples below), in particular she failed to include both the variable and the target population in her post-test descriptions and she used bullet points to note features rather than descriptive sentences.

Student A: post-test example – 7 features

> -Right skewed (9) and unimodal (10) with a tail to the right. -Peaks at approx. 20 (26). -Is approx symmetrical between 10–40 (13). -Is tightly grouped between 20–50 (21). [Shape sketched (7) and values (4) included.]

Student A: homework example – 13 features

> The distribution of the NZ (1) household debt (2) is right skewed (9) and unimodal (10). The debts peak (26) at approximately $10,000. The debts range from $0–$200,000 (18). Between $0–$80,000, the debts are approximately symmetrical, where they are tightly grouped (13, 21). There is a short tail from $100,000 to $200,000 (19). The middle household debt (24) is approximately $50,000. [Shape sketched (7), units – $ (3) and values (4) included.]

The post-test descriptions for student B were of a similar quality to her homework tasks. All of both students' descriptions in the homework tasks would have scored at the EA-5 level.

Between both students the homework descriptions used 16 of the possible 28 features. The additional features that were in the homework descriptions, but not in the post-test were descriptions of intervals, both the tail (19) and other groupings (20), and one mention of where the majority of the data (11) was, a feature also mentioned in the pre-test. Two of the features not mentioned in the homework description (or in the post-test as previously noted) were the mean (25) and mode (27) both of which had been mentioned in the pre-test. The remaining 10 features that were not mentioned in any of the descriptions analysed for this paper were: hypothesis and prediction (8), highest and/or lowest values (12), gaps (14), outliers (15), range as a value (16), inter-quartile range (17), majority (22) relative frequency (23), modal groups (28) and describing the statistical feature in context (5).

In addition, as the homework descriptions were analysed for the features used, it was noticed that both students wrote their descriptions with what appeared to be a very strong pattern. The actual features that were described tended to be the same features across all of their descriptions and in addition the way the description was written seemed to be similar. It appears that the students have developed their own sense of what a description should have and have written their descriptions following their own strong pattern. Two examples, one for each student, are given below and used to describe each student's pattern.

Student A – homework example

> The distribution of the weight (2) of these Great Spotted Kiwi (1) is approximately symmetrical (9) and bimodal (10). The weights of the kiwis peak (26) at 2.05 kg and at 3.3 kg (there are two peaks). There seems to be two groups of weights (20) – one from 2–2.5 kg and the other from 2.7–3.6 kg.

> The weights range (18) from 1.5–4.25 kg. The middle weight (24) is approximately 2.75 kg. I wonder if the bimodality is due to gender differences (6).

Student A starts with the introductory sentence, including the target population (1), variable (2), and overall shape with both symmetry (9) and modality (10) described. Then usually student A describes other shape aspects across sub-groups (13) and any groupings (19, 20) especially in skewed and/or bimodal graphs. This is generally followed up by range as an interval (18), peak (26) and median (24), but these follow no strict pattern, often being mixed from one description to the next. Student A also twice included an explanation for bimodality of a graph (6).

Student B – homework example

> The distribution of attendance (½ days) (2) for these year 9–13 students (1) is approx left skew (9). The attendance peaks (26) at 95%. The % of ½ days attendance range from 20%–100% (18). The middle % (24) is approx 90%. The % half days attendance are tightly grouped from 85%–100% (21). It is approx. symmetric in this range (13). There is a long tail to the left from 80%–20% (19).

Student B also has the introductory sentence with target population (1), variable (2), overall shape with both symmetry (9) and modality (10) described first. Then student B usually describes the peak (26), range as an interval (18), middle value (24) and then any groupings (13, 19, 20), especially in skewed and/or bimodal graphs.

CONCLUSION

The first question for this paper was to what extent are two year 10 students using the different distribution features in their pre- and post-test and homework responses? The two students used 18 of the 28 features listed in the description framework at some stage when describing distributions in either their pre- and post-test responses or in their homework responses. The features that they used were appropriate for the examples given in both the tests and homework. Some of the features that the students did not use were not present in the graphs given. To improve students' use of these features additional graphs with the particular features need to be sourced and then used in the teaching and learning program. For example, gaps and outliers were specific features that were missing from the graphs provided. Some of the other features, such as range, interpretation, highest and lowest values, were present in the graphs given and their absence suggests that they were not used or highlighted in the teaching and learning sequence. That is, the teacher did not select them as features when she was leading the writing, or when she was actively reflecting on student work from the homework responses. Hence for students to describe and reason from distributions the learning trajectories need to pay attention to providing experiences with a range of distributional shapes and contexts (Bakker, 2004b).

The second question for this paper was how are these two year 10 students using the different distribution features in their pre- and post-test and homework responses? Both students appear to have a pattern to the way that they write their descriptions with student B being more structured than student A. The pattern refers first to the context with an overall impression of the plots, focuses on the details and then wonders about explanations for particular features, which is a similar pattern to that described by Pfannkuch et al. (2010) as a way of looking at and interpreting plots. Hence their descriptions seem to be reflecting how a statistician would reason from plots.

The distribution framework is a tool that teachers can use to support their selection of graph examples to use with students when teaching and assessing distribution descriptions and to ensure that the all the features are covered in the different examples given. Once the range of features is attended to in the learning and teaching, the distribution framework can also be used to diagnose which of the features the students are not attending to in their descriptions. Researchers can also use the distribution framework as a tool for designing assessment instruments and for investigating and developing students' reasoning processes about distribution.

REFERENCES

Arnold, P. (2013). *Statistical investigative questions: An enquiry into posing and answering investigative questions from exisiting data*. Doctoral thesis. The University of Auckland, New Zealand.

Arnold, P., & Pfannkuch, M. (2012). The language of shape. *Proceedings of the 12th International Congress on Mathematical Education (ICME-12, July, 2012), Seoul, South Korea* (pp. 2446–2455). Seoul, South Korea: ICME-12.

Bakker, A. (2004a). *Design research in statistics education: On symbolizing and computer tools*. Utrecht, The Netherlands: Freudenthal Institute.

Bakker, A. (2004b). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64–83.

Ben-Zvi, D., & Amir, Y. (2005). How do primary school students begin to reason about distributions? In K. Makar (Ed.), *Reasoning about distribution: A collection of research studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-4, July, 2005), Auckland, New Zealand* (pp. 1–22). Brisbane, Australia: University of Queensland.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics, 45*, 35–65. doi: 10.1023/A:1013809201228

Ben-Zvi, D., Gil, E., & Apel, N. (2007). What is hidden beyond the data? Helping young students to reason and argue about some wider universe. In D. Pratt & J. Ainley (Eds.), *Reasoning about statistical inference: Innovative ways of connecting chance and data. Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5, August, 2007), University of Warwick, England* (pp. 1–26). Warwick, England: University of Warwick.

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

delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In K. Makar (Ed.), *Reasoning about distribution: A collection of research studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-4, July, 2005), Auckland, New Zealand* (pp. 1–17). Brisbane, Australia: University of Queensland.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124–158.

Hancock, C., Kaput, J., & Goldsmith, L. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 2*(3), 337–364.

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2004). *Data seen through different lenses*. (Unpublished manuscript.) Scientific Reasoning Research Institute, University of Massachusetts, Amherst, MA, and TERC, Cambridge, MA.

McClain, K., & Cobb, P. (2001). Supporting students' ability to reason about data. *Educational Studies in Mathematics, 45*, 103–129. doi: 10.1023/A:1013874514650

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal, 5*(2), 27–45.

Pfannkuch, M., Arnold, P., & Wild, C. J. (2011). *Statistics: It's reasoning not calculating.* (Summary research report on Building students' inferential reasoning: Levels 5 and 6) Retrieved from http://www.tlri.org.nz/tlri-research/research-completed/school-sector/building-students-inferential-reasoning-statistics

Pfannkuch, M., Regan, M., Wild, C. J., & Horton, N. (2010). Telling data stories: essential dialogues for comparative reasoning. *Journal of Statistics Education*, *18*(1), 1–38.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: from a variation perspective. *Statistics Education Research Journal*, *5*(2), 46–68.

Roth, W.-M. (2005). *Doing qualitative research: Praxis of method*. Rotterdam, The Netherlands: Sense.

Watson, J. (2005). Developing an awareness of distribution. In K. Makar (Ed.), *Reasoning about distribution: A collection of research studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-4, July, 2005), Auckland, New Zealand* (pp. 1–33). Brisbane, Australia: University of Queensland.

Wild, C. J. (2006). The concept of distribution. *Statistics Education Research Journal, 5*(2), 10–26.