SYMBOLIZING DATA INTO A 'BUMP'

Arthur Bakker, Freudenthal Institute,
Utrecht University, Utrecht, the Netherlands
email A.Bakker@fi.uu.nl

Abstract

In this paper we analyze how concepts and symbolizations co-develop in the case of statistical data analysis. The focus is on the development of distribution, which ranges from a very concrete intuitive understanding to formal mathematical definitions. Examples from teaching experiments with 11 to 12 year-old students illustrate how their concept of distribution develops in relation to what the graphs they use and make mean for them. In particular we discuss an episode in which a student symbolizes data into a so-called 'bump' and we give examples of how other students reason with this 'bump' in connection to distribution.

Introduction on Symbolizing

Symbolizing as a field of research has been receiving more and more interest within the community of mathematics educators (Cobb et al. 2000). Our point of departure is that the students' way of symbolizing and what these symbolizations come to signify develop in a dialectical way. The learning-teaching process is organized in such a way that the conceptual development benefits from the development and use of symbols, and vice versa. In teaching experiments on statistical data analysis we aimed for the gradual emergence of the multifaceted concept of distribution rather than a collection of loosely related concepts and graphs. In this paper we analyze how symbolizations and their meaning co-evolve in the case of distribution and graphs of data sets.

The Concept of Distribution

Basically, the notion of distribution refers to how data are distributed in a space of possible values. Mathematically seen, distribution could be defined as a frequency or density function. In this paper we use the term 'distribution' for the whole range from a very concrete, intuitive level to the statistical concept. The concept of distribution is tightly connected to many other statistical concepts such as frequency, skewedness, spread, and even mean and median. It is also highly interwoven with certain visual images that 'show' distributions, the most famous one being the bell-shaped curve of the normal distribution.

In statistical data analysis we are not very interested in the individual cases; instead we focus on group characteristics such as center and spread, or skewedness of data. Even

if we are going to use the mean or median we have to take the whole distribution into account (Zawojewski & Shaughnessy 2000). For example, if there are many outliers or if the distribution is very skewed we probably will not use the mean. From the numerical data it is hard to see how they are distributed, so we need to look at the shape of the data.

There are two other reasons we focus on distribution. First, students tend to see data as individual cases instead of attributes or a value of a variable (Hancock et al. 1992). That is why they find it difficult to see group characteristics. Focusing on the shape of the data is one way of dealing with this problem. Second, we wanted students to reason proportionally instead of absolutely with parts of the graph or the data. By proportional reasoning we mean reasoning with proportions as opposed to absolute numbers. If students, in comparing parts of two samples with different size, reason with absolute numbers then we call it absolute reasoning, whereas if they reason with proportions we call it proportional reasoning (Cobb (1999) calls this multiplicative reasoning). If students focus on the shape of distributions they are supported in reasoning more globally with groups and are maybe led away from absolute numbers. A helpful tool for this purpose is the box plot since it shows proportions, for instance where the middle 50% of the data is, without showing individual data points.

For statisticians distribution has a clear experiential meaning. For 11-year-old students, however, distribution initially is not on the horizon. Still, young learners can deal, for example, with questions concerning the way data are distributed. They can solve problems that involve looking at how the data are spread out or bunched up. To foster this kind of reasoning demands that the tasks in statistical data analysis must have certain features. First, they have to conduct the students' reasoning to characteristics of the distribution, even if the students do not talk in terms of distributions. Second, these characteristics must be expressible both in terms of the context and with statistical concepts. Anticipating the next section we mention that just one problem concerning the life span of batteries initialized discussions around center, spread, outliers, majority in terms of the context. A sample of a good battery brand has a high mean and a sample of a reliable brand has small spread and few outliers (figure 1). Of course none of these notions were very precise yet, but still they formed a basis for the development of more formal concepts.

In solving such problems the students used so-called statistical minitools. These software tools have been designed for the teaching experiments of Cobb, Gravemeijer, and others (Cobb 1999; McClain et al. 2000). The minitools do not contain any ready-made conventional statistical graphs. Instead the students can structure the data in various ways they might use when they just have pencil and paper. Still, some grouping options are precursors to conventional graphs such as using equal interval width underpins the histogram and using four equal groups underpins the box plot (figure 2).

Value Bars Come to Signify Data

The first step of symbolizing was inscribing data as case value bars, which offers a visual way of dealing with the data (figure 1). The horizontal bars are motivated by a sense of linearity that many variables have. The first data set analyzed by the students concerned the life span in hours of two battery brands. The task was to decide which brand was best and to report an analysis to the Consumer Reports.

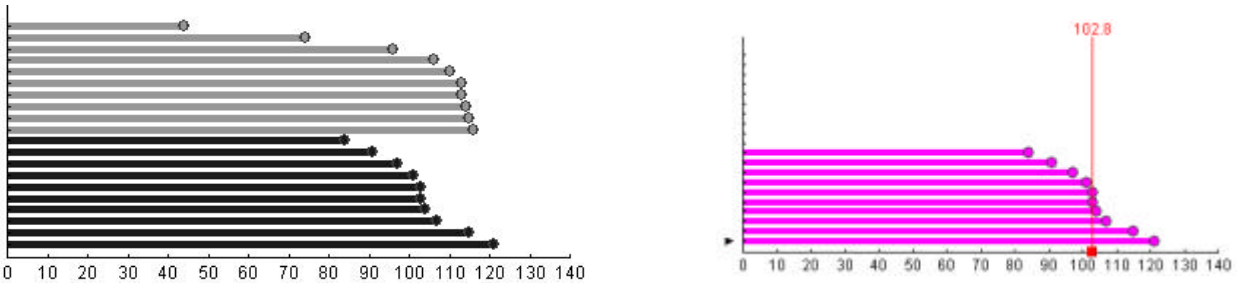


Figure 1. *Value bar graphs in minitool 1. 1a: Life spans of two brands of batteries in hours, in the minitool indicated by two different colors. The upper ten and lower ten are of different brands. 1b: Visually finding the mean with the value bar.*

Already in this type of graph it is visible, for statisticians, that the distribution of the first brand is skewed and has outliers at the left; the second brand has smaller spread and has a symmetrical distribution. Here distribution is symbolized on a very concrete level. The students did not talk of distributions but they discussed the outliers of the first brand and the high maximum of the second brand. Some argued that the first brand was better since ‘it has more higher values’ and others opposed that the second brand ‘has less bad outliers’. The second brand was considered more reliable or predictable (compare this with the notion of consistency on which Cobb (1999) and Sfard (2000) report). Students were very well able to invent data sets that could be of a very good but unreliable brand or a bad brand with the same spread as one they had encountered before.

After a few lessons the students developed a visual way of estimating means by mentally cutting off ‘what was too much on the right side’ and ‘giving it to the left side’, as they express it (see figure 1b). Such an activity was made possible by the inscription of the case value bars. We conjecture that students would not have done this with numerical data or with dots in a dot plot. This supports our claim that every symbolization influences the way students see the data or the context problem. On the other hand, a suitable problem may give rise to developing a new symbolization that helps in answering a question. Also on this level there is a dialectical co-development of meaning of the context on the one hand and of the symbol on the other (cf. Meira 1995).

Dots Come to Signify Value Bars

As the preceding section shows, the students already developed a statistical language that was situated in the context of battery life spans and other problems. In thinking about such problems the students focused on the end points of the bars. These end points of the bars in minitool 1 were to collapse down onto a horizontal axis in minitool 2. The dot plot appears here as an image of a variable; the dots get a place in a space of possible values. This dot plot is also one step closer to the conventional graphs in which a unimodal distribution appears as a hill-shaped curve. As mentioned above some grouping options were close to conventional graphs such as histogram and box plot (figure 2), both helpful tools in describing distribution.

In analyzing data with this second minitool, students further developed their statistical language. Since we wanted students to view data sets as a whole with certain characteristics we hoped that they would start looking at the shape of the data in minitool 2. Unfortunately they did not talk of hills as students in other experiments did (Cobb 1999); they only talked of majorities and still reasoned additively in some cases.

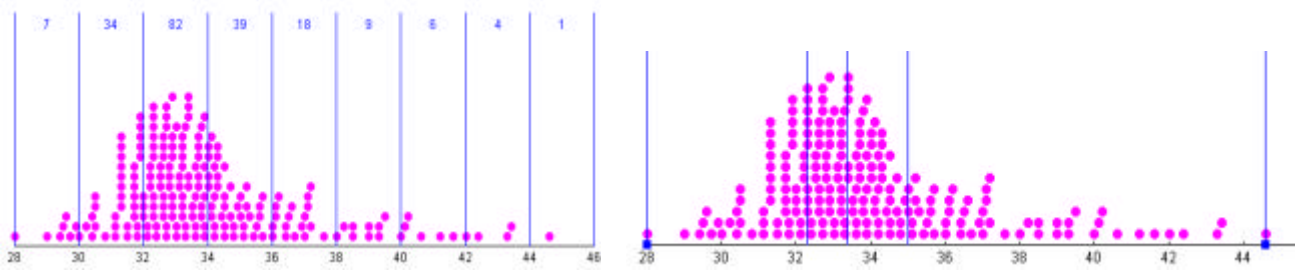


Figure 2. *Dot plots in minitool 2. 2a: The option of equal interval width is used to organize the data, underpinning the histogram. 2b: The option of four equal groups underpins the box plot.*

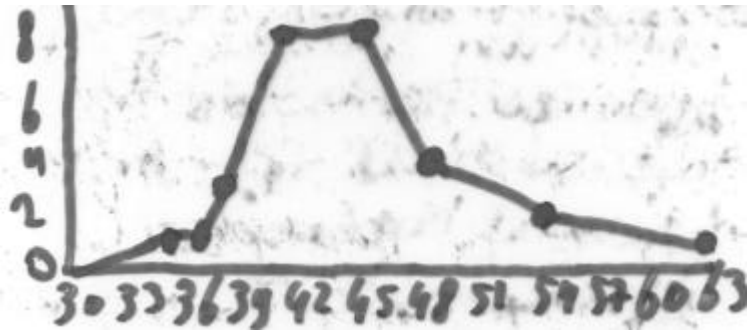
Symbolizing Data into a 'Bump'

We tried another route that turned out to be more promising. The basic ideas of symbolizing and guided reinvention (e.g. Gravemeijer 1994) suggest that students also should create their own graphs. For the eleventh lesson we therefore asked them to represent their weight and height data in a graph that would be clear for a particular purpose: a balloon rider had to decide how many seventh-grade students could join a balloon ride and she did not just wanted to know the mean.

The students came up with many different graphs resembling minitool 1 and minitool 2, but also a scatter plot of weight and height, and one graph we will discuss in more detail. For the designed learning process we focused on graphs that could help students in seeing a data set as a whole, or in other words, could help them in constructing distribution as an object-like entity which they could reason with. This was why

Michiel's graph (figure 3a) was discussed extensively. He explained his graph as follows.

Michiel: Look, you have roughly, averagely speaking, how many students had that weight and there I have put a dot. And then I have left [y-axis] the number of students. There is one student who weighs about 35 [kg], and there is one who weighs 36, and two who weigh 38 roughly. And then I have put, yeah/
Teacher: Have just put dots.
He then explained in more detail what the dots stood for. The dot above 48, for example, signifies that four students had weights around 48 kg. After discussing other graphs the teacher asked the following question.



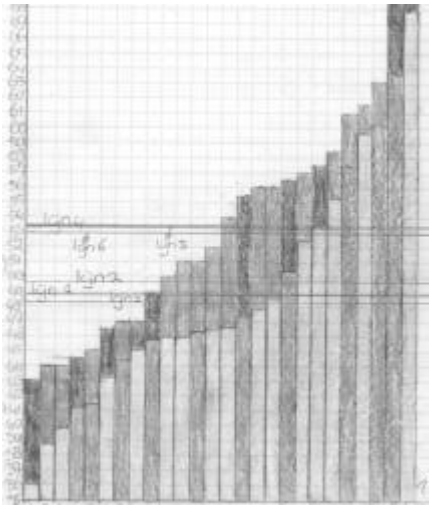


Figure 3a and 3b. *Michiel and Elleke's graphs of weight data. Elleke's includes height represented by the darker and higher bars. Her graph is a value bar graph.*

Teacher: What can you easily see in this graph [of Michiel]?
Laila: Well, that the average, that most students in the class, um, well, are between 39 and, well, 48.
Teacher: Yes, here you can see at once which weight most students in this class roughly have, what here is about the biggest group. Just because you see this bump here. We lost the bump in Elleke's graph.

It is the teacher who uses the term 'bump' for the first time. Later in the discussion one student explained where the bump in Elleke's graph was.

Nadia: The difference between … they stand from small to tall, so the bump, that is where the things, where the bars are the closest to one another.
Teacher: What do you mean, where the bars are closest?
Nadia: The difference, the ends [of the bars], do not differ so much with the next one.
Another student commented on this.

Eva: If you look well, then you see that almost in the middle, there it is straight almost and uh, yeah that/ [teacher points at the horizontal part in Elleke's graph].
Teacher: And that is what you [Nadia] also said, uh, they are close together and here they are bunched up, as far as height or weight is concerned.
Eva: And that is also that bump.
From these excerpts and further analysis of the episode (Bakker 2001) it became clear that these students did not just rely on visual aspects of the bump. They were able to relate the different graphs to one another by thinking of what the bars and dots signified. From the analysis of this lesson the question remained whether the bump just signified the majority for the students or that it signified a characteristic of the whole distribution.

Reasoning with the 'Bump'

From consequent lessons we inferred that the bump not just signified the majority. We give a few examples of how students reasoned with the bump. What is interesting in this respect is that many students used the term 'bump' even for the straight part of value bar graphs for where most data were. It became clear that they related this visual characteristic of the symbol also to statistical concepts such as outliers and sample size.

Laila: But then you see the bump here, let's say.
Ilona: This is the bump [pointing at the straight vertical part of the lower ten bars, like in figure 1b].
Researcher: Where is that bump? Is it where you put that red line [the value bar]?
Laila: Yes, we used that value bar for it (...) to indicate it, indicate the bump. If you look at green [the upper ten], then you see that it lies further, the bump. So we think that green is better, because the bump is further.
Here the bump seems to have become a reasoning tool.
One question in a class discussion was what a graph of eighth-graders' weight would look like. Some of the answers follow.
Luuk: I think about the same, but another size, other numbers.
Guyonne: The bump would be more to the right.
Teacher: What would it mean for the box plots?
Michiel: Also moves to the right. That bump in the middle is in fact just the box plot, that moves more to the right.
Turning to a different question, the researcher (being the author) asked the class how the graph would change if not just their own class but all seventh–graders in the province were measured. He was curious if the bump only signified the majority or that it was also linked to outliers and sample size. Earlier in the class discussion Elleke had mentioned that with more students one has more chance for outliers.

Elleke: Then there would come a little more to the left and a little more to the right. Then the bump would become a little wider, I think.
Researcher: Is there anybody who does not agree?
Michiel: Yes, if there are more children, then the average, so the most, that also becomes more. So the bump stays just the same.
Albertine: I think that the number of children becomes more and that the bump stays the same.
Nadia: I think that if there are less children, you have more chance for outliers. Maybe some are very thin and some very heavy or so. But I think that it stays roughly the same.
A few students were able to see in figure 1a which distribution was 'normal'—defined in an informal sense—and which was skewed.

Albertine: Oh, that is normal (...).
Nadia: That hill.
Albertine: And skewed if like here the hill is here [the upper ten bars].

Conclusions

In the process of symbolizing data were first inscribed as value bars in the first minitool. The end points of the bars collapsed down onto an axis and formed a dot plot. It was shown that every symbolization has its advantages. The bars made it reasonable to the students to find means in a visual way and the dots made it easier to structure the data in other helpful ways. Finally, one of the students' graphs led to interesting discussions about bumps. At this stage, the learners were able to reason in a more global way without focusing on individual data points and they also argued in a more multiplicative way.

As we demonstrated, focusing on the concept of distribution, being a multifaceted notion, has the advantage that it is strongly related to almost all other statistical concepts. In this way we could help students to gradually build up their understanding of these concepts in close relation to one another.

We showed how students came to construct an intuitive understanding of distribution in close relation to how they come to signify meaning to a series of graphs. As the examples illustrate the students related statistical concepts to characteristics of the graphs. It is clear that the development of a concept, in this case distribution, cannot be separated from the development of the symbols.

Acknowledgements

I thank Mieke Abels for teaching in these experiments. The analysis carried out in this paper was supported by NWO, the Dutch National Science Foundation, under grant no. 575-36-03B. The opinions expressed do not necessarily reflect the views of the

Foundation. The research project of the researcher is part of a larger project called ‘Mathematics and IT’, which is coordinated by Prof. K.P.E. Gravemeijer.

References

Bakker, A. (2001). Symbolizing data into a ‘bump’; Different theories of symbolizing on one episode. Manuscript, Utrecht.

Cobb, P. (1999). Individual and Collective Mathematical Development: The Case of Statistical Data Analysis. *Mathematical Thinking and Learning,* 1(1).

Cobb, P., E. Yackel, & K. McClain (eds.) (2000). *Symbolizing and Communicating in Mathematics Classrooms; Perspectives on Discourse, Tools, and Instructional Design.* Lawrence Erlbaum Associates, Publishers: Mahwah, New Jersey/London.

Gravemeijer, K.P.E. (1994). *Developing Realistic Mathematics Education***,** Freudenthal Institute. CD Beta Press.

Hancock, C., J.J. Kaput & L.T. Goldsmith (1992). Authentic Inquiry with data: critical barriers to classroom implementation. *Educational Psychologist*, *27*(3), 337-364.

McClain, K., M. McGatha, & L.L. Hodge (2000). Improving Data Analysis Through Discourse. *Mathematics Teaching in the Middle School, 5* (8), 548-553.

Meira, L. (1995). Microevolution of mathematical representations in children’s activity. *Cognition and Instruction*, *13*(2), 269-313.

Sfard, A. (2000). Steering (Dis)Course Between Metaphors and Rigor: Using Focal Analysis to Investigate an Emergence of Mathematical Objects. *Journal for Research in Mathematics Education, 31*(3), 296-327.

Zawojewski, J.S. & Shaughnessy, J.M. (2000). Mean and Median: are They Really so Easy? *Mathematics Teaching in the Middle School, 5* (7), 436-440.