Drawing Connections from Modelling Distribution to Data-centric Distribution

Theodosia Prodromou
University of New England, Armidale
theodosia.prodromou@une.edu.au

Abstract: *Drawing connections between theoretical frequencies and observed relative frequencies is advocated by various researchers. Little is known, however, about the process by which these connections are developed. The focus of this paper is on students' (aged 14 to 15 years) constructions of meanings about the intentionality interpretation when co-ordinating the two perspectives on distribution. Extracts from two research studies illustrates students' construction of the intentionality route from the model (modelling distribution) to the data (data-centric distribution) through their attempts to transform, the specific modelling distribution, and observing how that changes a graph/histogram of the actual outcomes, by using on screen control mechanisms to change the way that the computer generates the data within a carefully design computer simulation. In many respects their interpretations were based on similar characteristics to those of experts.*

Keywords: modelling, distribution, simulation, randomness

1. Introduction

The Australian Curriculum and Reporting Authority K-10 (ACARA, 2010) recognizes that the twenty-first century world is information driven, and through Statistics and Probability students can make informed judgments about events involving chance. The role of probability as a central component for statistical investigations has engendered the need for "probability literacy" to deal with a variety of real-world situations that encompass interpretation or generation of probabilistic messages as well as decision-making.

A proposed view of probability literacy illustrates a "scope of probability at the school level to reflect the study of random events, the development of appropriate probabilistic intuitions, a basic understanding of language and simple events, an appreciation of distribution, and the addressing of misconceptions" (Watson, 2006, p. 127-128). This scope can be seen as consistent with the tendency to place less emphasis on the knowledge pertaining to theoretical probability when addressing important issues relevant to teaching data-based statistics (Moore, 1997). Using a computer simulation of a basketball player, and observing histograms generated to describe the performance of the simulated basketball player, students (aged 14 and 15 years) were observed to develop two separate interpretations to explain the connection between the two distributions. Continuous classroom-based investigations based on frequencies may reinforce the building of an appreciation of a frequentist approach to probability when performing trials and comparing favourable outcomes to total outcomes.

Such an emphasis on the frequentist approach, including the increasing interest in statistics and the growing expectations around what is expected from students when handling data, gives statistics the bulk of attention in current curricula, while the theoretical models based on the formal sample space and the theoretical probabilities are not given the appropriate attention. For example, Watson claims (2006) claims that it is not always appropriate to introduce an experiment to calculate relative frequencies before suggesting a theoretical model based on the possible outcomes of a sample space.

Ultimately, introducing the environment of theoretical probability where data will be collected should not be left behind or dismissed. Hands-on simulations and simulation software provide students with opportunities to explore the

nature of the sample space and probability distributions. In this framework, for instance, a probability distribution of some discernible characteristics has the status of a model of the data that describes what one could expect to see if many samples would be collected from a population, enabling us to compare data from a real observation of this population with a theoretical distribution.

In order to better understand how a probability distribution has a status of a model that generates data, the researcher has developed a simulation and analyses the students' articulations, in which they use "situated abstractions," (Noss & Hoyles, 1998) to guide our attention to important elements in the students' thinking processes. Hence, this article examines how students perceive the theoretical distribution as a model that generates data and how they build connections from the theoretical distribution towards the data distribution.

1.1. Basketball Simulation

Prodromou (2008) built a computer-based simulation of a *BasketBall* simulation player attempting to make a basket (Figure 1). Underlying the simulation were two mechanisms for generating the trajectories of the balls following Newton's Laws of Motion; one was fully deterministic; the other used a probabilistic model that incorporated variation in the trajectories. The interface was designed in such a way that the data-centric perspective was presented graphically as a set of data about the trajectories and success of shots at the basket, and the modelling perspective was presented as the probability distribution that generated the varying trajectories of the balls. Prodromou's research was interested in investigating whether and how students co-ordinated the experimental outcomes (data-centric perspective) and the theoretical outcomes (modelling perspective).

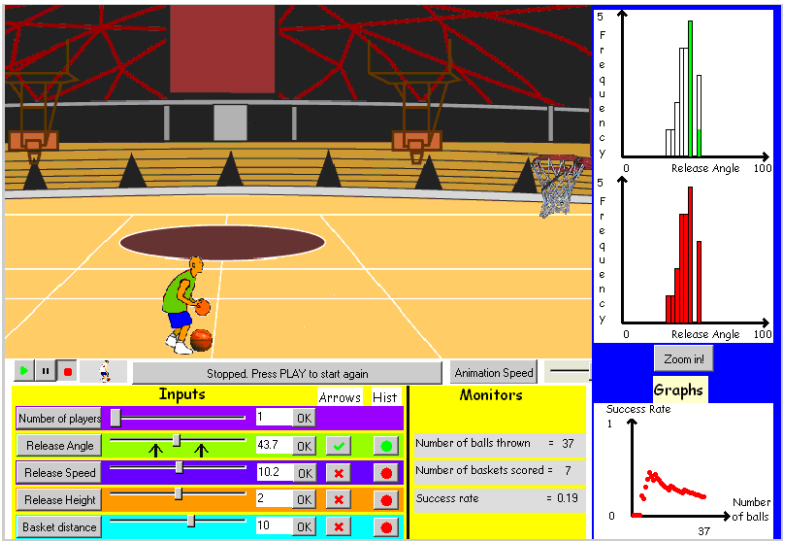

Figure 1.

The player has successfully thrown the ball into the basket. The release angle, speed, height and distance can all be varied using the sliders or by entering the data directly. Once the play button has been pressed, the player continues to throw with the given parameters until the pause or stop button is pressed. The graph in the bottom right hand corner traces the relative frequency of successful throws.

The students were able to control the model that generated the data, and through this control they had access to information about the more general modelling perspective on distribution. In particular, students were instructed to open a dialogue box that showed the distribution of values for a variable from which the computer would randomly choose, given the student's settings of the handle of the slider and the arrows (Figure 2). The students were able to move either the arrows or the handle on the slider and observe corresponding changes in the graphical representation of the modelling perspective on distribution.

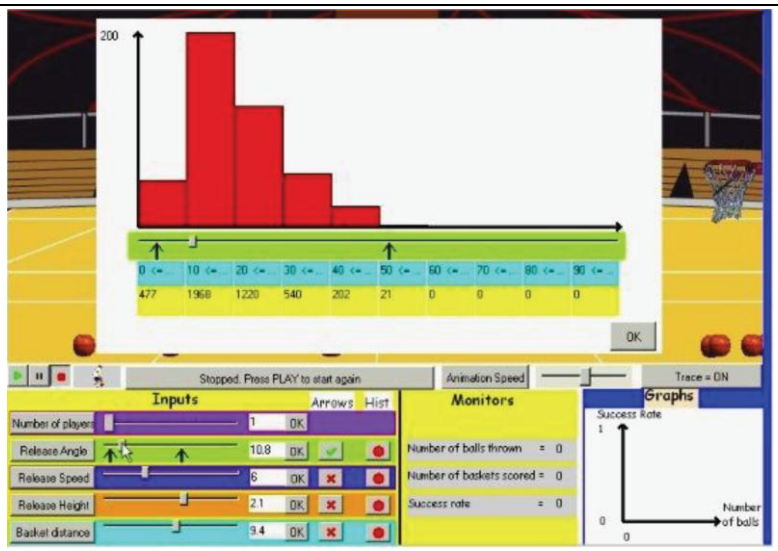

Figure 2. The students altered the modelling perspective directly by moving the arrows or the handle on the slider.

After a few minutes of the simulation play, students were asked to observe the impact of their actions on the outcome histograms when manipulating the arrows and the handle of the slider. The students compared the graph of the modelling perspective to the graph that plots the data generated by the computer (i.e., the data-centric perspective, on the right of the screen), and attempted to make interpretations of the relationship between the perspectives (Figure 3).

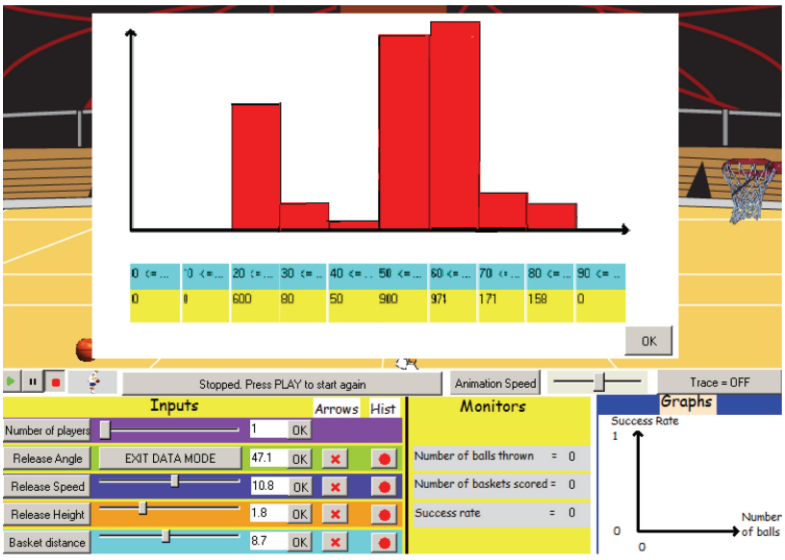

Figure 3. The students altered the modelling perspective directly by setting numerical values associated with each possible outcome for a given variable.

2. Method

The data come from a research study with eight pairs of 14 to 15-year-old students in a UK secondary school. Each pair of students participated in the completion of tasks in four 45-minute sessions.

In these sessions the students spent extensive time working in pairs. The researcher (res) interacted continuously with the pairs of students during the pair work phase in order to probe the reasons that might explain their thinking. The data collected included audio recordings of each pair's voices and video recordings of the screen output on the computer activity using Camtasia software.

The recordings of the eight pairs of students were fully transcribed shortly after the end of each session. Interpretative case analyses were developed which criticise why and how students' constructions of meanings developed. These

analyses became the main focus for subsequent analysis and triggered further phases of progressive focusing (Robson, 1993) to identify key foci for ensuing study. Important similarities and differences between the interpretative case analyses were then identified by constant comparisons (Glaser, 1978) of these eight interpretive case analyses.

I report in this paper on one key focus, the different ways in which the students perceived the modelling distribution as a model that generates data and built connections from the modelling distribution towards the data distribution while exploring the generation of data within the BasketBall simulation. This article focuses on two pairs of students, James and Anna; Matthew and Emily.

3. Findings

When Anna and James looked at both the modelling distribution and the data-centric distribution, they recognized that the two perspectives on distribution were showing different things. They were then asked to compare the shapes of the graphical representations of the modelling and data distribution (Figure 4).

1. Anna: Ehm … They are similar, but not the same, because they both have got the tallest bar and then two shorter …
2. Res: How can they can be similar but at the same time tell different things.
3. James: Because, they've got the tallest in the middle, the second tallest on the left and then the second tallest on the right … on all the graphs … on all the graphs …

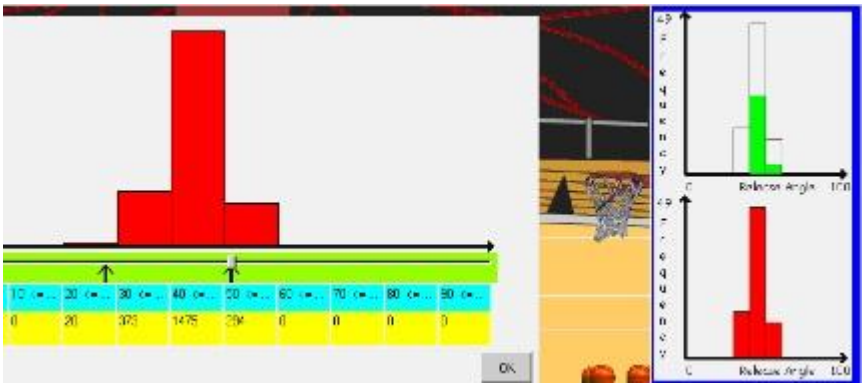

Figure 4. The modelling distribution (graph on the left) is showing the way in which the values of the corresponding variable were chosen. The graphs on the right (data-centric perspective on distribution) plot the data generated by the computer, divided into successful throws (top) and attempts (bottom).

Their articulated reaction at this stage was restricted merely to comparing isolated bars of the data-centric distribution to their corresponding bars of the modelling distribution. This type of reaction was followed by considering the shape of a distribution as if it were an accumulation of just a few isolated bars: "the tallest in the middle, the second tallest on the left and then the third tallest on the right." A few minutes later, I suggested that they look once more at the histograms.

4. Res: You said earlier that the two graphs tell us different things.
5. James: Yes, but their shapes are becoming the same … Yeah, because that one is showing you what angle you selected (he was talking about the modelling distribution).
6. (I ask him to show me with the mouse.)
7. James: (points to the graph on the left) This is showing what you selected and then these angles … they are showing which one you are using.
8. Res: Yeah … but the shape as we can see is gradually becoming the same … Is it a coincidence?
9. James: No. It's meant to do that.

After allowing the simulation to run for a while, James noticed that the shapes of the two graphs were becoming more alike. He tried to associate the two distributions by looking closely at the effect on the animation of throwing more balls.

It is interesting when James says 'selected' rather than say 'could select,' he seems to resort to a human agency. The human agent can either be the researcher or the basketball player (see Prodromou, 2008; Prodromou & Pratt, 2009).

He distinguished the model (modelling distribution) from the real data (data-centric perspective), but the two different perspectives on distribution were not as yet well-formed into his mind as complementary or related concepts.

It would appear that James recognised that the histogram of frequency of successes against angle materialises from the modelling distribution due to the students' changes to the model when they varied the slider to change the relative frequencies in the modelling distribution.

James and Anna did not comment on the differences and similarities of the two perspectives of distribution and at this point I wished to probe into their intuitive understanding and informal reasoning about basic features (probability, shape, spread and skewness, percentage of central cases) of the data-centric distribution and the modelling distribution.

10. James: Similarities are that the tallest one, the tallest bar on the middle. The second tallest bar is on the left, and … the last one is on the right … ehm … the difference, is this one … it tells you (pointing to the bar on the left) … this tells you what angle you selected, which one he gonna use the most and this one (the graph on the right) tells you how he used all together in each throw, and this shows you the success rate of all the throws.
11. Res: Are their shapes the same?
12. James: Yeah.
13. Res: So, the basketball player was meant to play as he did.
14. James: Because, we've told him to. Well, it's meant to play between those lines, but on that slider.

When the students talked about the similarity of the two perspectives of distribution, they did so in terms of the relationship between the heights of the bars, comparing, thus, slices of data.

In contrast, when they referred to the differences, they appealed not to the specific data, which was self-evidently different from bar to bar, but rather to the underlying role of the two distributions. They referred to modelling as what was intended but to the data distribution as what actually happened. James, however, talked comfortably about variation, that was a characteristic of the data-centric perspective. At this point, the vicissitudes of randomness and the lack of a strong sense of the probabilistic mechanism had prevented James and Anna from talking explicitly about chance and randomness. I think this can be attributed to their lack of a clear probabilistic-type language for talking about randomness and probability. Although they recognised that the shapes of the two graphs were the same when viewed holistically, James and Anna showed scant argumentative capacity in making a judgement about how well the model fits the data. Only at the end, James took into consideration the overall spread.

I now wish to focus on the work of Emily and Matthew in which distribution was perceived in terms of intention. After Matthew and Emily had spent 15 minutes moving either the arrows or the handle on the slider and observed the impact of their actions on the graphical representation of the modelling perspective on distribution, they were faced with the graphical representations of the modelling and data-centric distribution (Figure 5).

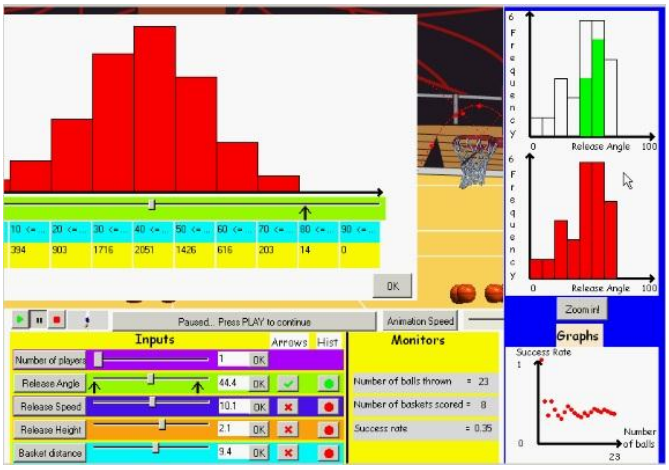

Figure 5. Emily and Matthew looked at the modelling distribution (graph on the left) and the data-centric perspective on distribution (graph

on the right).

15. Res: What is the graph on the left hand side about?
16. Matthew: It's about the chances of success you should have (graph on the left)…That's how it should turn out … (graph on the right).
17. Emily: But it's about release angles.
18. Matthew: It's showing which angles … would be chosen… (graph on the left) Yes, that's showing which angle is the most likely angle you score it … (graph on the right).

Matthew seemed to know the distinction between the data-centric perspective on distribution, which identifies distribution as an aggregated set of actual outputs, and the modelling perspective on distribution, which views distribution as a set of possible outcomes and associated probabilities. According to the articulations of Matthew, there was certainly a notion which performs the function of intentionality, and this is precisely portrayed by his situated abstraction "That's how it should turn out …" that the data distribution is anticipated to equal the modelling distribution. Matthew continued to pay attention to the emerging data, by considering the data-centric perspective on distribution, and simultaneously used information from the modelling perspective from which the data was generated (sampled).

19. Res: What are the graphs on the right about?
20. Matthew: That are showing … those angles …yeah (very excited) ….they are the same, because if you look at them, the success rate on these two is the most …
21. Res: Yeah … show us.
22. Matthew: What I was saying is: These are the top two (pointing to the highest bars of the graph on the left) and I reckon we haven't any of them (pointing to the lowest bars of the graph on the left) and then these two (pointing to the highest bars of the graph on the right) and these two (pointing to the tallest bars of the graph on the left) and then we have shots on all of these (pointing to the bars 0−40 of the graph on the left hand side)…but we only succeeded in these two.
23. I say: Do you mean from 30 to 50?
24. Matthew: Yeah.
25. Res: What is the tallest bar? (pointing to the highest bar), Is it from 40 to 50?
26. Matthew: Yeah.
27. Res: What is the tallest bar of the graph on the left hand side?
28. Emily: 40 to 50.

Emily and Matthew expected a hidden order of some kind, based on the numbers presented within each bar of the modelling perspective on distribution, on their selection and arrangement on the data-centric perspective on distribution. This systematic reaction, which we observed principally in all the other case studies, was really not incomprehensible for students because they had already possessed the deductive order. They anticipated the data to be chosen mainly from the highest bars of the modelling distribution. At that stage, the generation of the data-centric perspective on distribution from the modelling distribution appeared to be precisely the contrary of such an order. They tried subsequently to reason about the mechanism which was masked behind the fortuitous generation of data from the modelling perspective and resisted any deduction.

29. Res: Yeah … look now at the (top) graph on the right hand side … it's completely different.
30. Emily: The angles are going between the arrows and it's gonna use all the different angles between the arrows.
31. Res: Do you remember earlier these two bars (pointing to 20−30, 30−40 bars)?
32. Matthew: I know it grows, because you are throwing more balls.
33. Res: Yes, they had the same height. I am talking about the bars from 30−40, and 40−50.
34. Matthew: I know they are different, because you've got more chance of scoring it and these 3 here (pointing to the bars between 30−60 of the graph on the left hand side). So, those bars are higher (pointing to the same bars of

the (top) graph on the right hand side), because the computer wanted to select these … (pointing to the bars 30−60 of the graph on the left hand side). But because it still succeeded in this one…they are taken into account.

Matthew recognized which values from the modelling distribution had more chance to be selected from the modelling perspective of distribution. Indeed, he very nicely articulated that "you've got more chance of scoring it and these 3 here (pointing to the bars between 30−60 of the graph on the left hand side). So, those bars are higher (pointing to the same bars of the graph [top] on the right hand side)." He attributed the selections of the emerging data as apparent in the data-centric perspective on distribution, to randomness which, in turn, lies within the confines of the *BasketBall* simulation (computer). The computer, therefore, is acting as agent of randomness. In fact, this transition of agency seems to reflect a move towards modelling.

This is articulated in terms of stochastic intentionality (line 34).

4. Summary

When students manipulated the ICPs, we saw that both the pairs, Anna/James and Emily/Matthew, appeared to perceive the modelling distribution as the intended outcome and the data-centric distribution as the actual outcome. Anna and James, however, had a sense of a general intention (I_G) when they talked about variation. Their articulations were explicitly characterised by the absence of a strong sense of the probabilistic mechanism and as evidenced, there was not a progressive articulation of intuitive relationships leading gradually to a clear probabilistic-type language for talking about randomness and probability.

Their reasoning about intentionality remains insufficiently clear. We can have at least two different possible interpretations: a) the intention is simply an expression of the pre-programmed deterministic nature of the computer – at least in their experience, or b) intentions are reflected in the actions of a modelling builder.

Emily and Matthew gave evidence of a sense of a stochastic intention (I_{st}; lines 29-34). Indeed, they seemed to see the modelling distribution as the intended outcome, progressively generating the data-centric distribution. It can be concluded from their various references to chance that this perception of intention is probabilistic, and underpins the idea that the modelling distribution precedes, indeed generates, the data-centric distribution. Hence, Emily and Matthew refer to the computer 'wanting' to throw the ball at various angles, a sentiment which we characterise as the situated abstraction, "the more the computer wants to throw at a particular angle, the higher is that angle's bar".

When students constructed a general Intention interpretation (I_G), they appeared to recognize a movement from the modelling distribution towards the data-centric distribution.

5. Conclusion

The intentionality route shows how some students perceived of the modelling distribution (MD) as the intended outcome and the data-centric distribution (DD) as the actual outcome, suggesting a connection being made, in which the modelling distribution in some sense generates the data. Intention does operate in that direction and not surprisingly general intention has the potential to become stochastic intention (I_{st}) when randomness becomes a part of the interpretation of the student.

When students attempted to make connections from the modelling distribution (MD) to data-centric distribution (DD), they typically tended to articulate simple causal explanations while observing the modelling distribution that generated the throws. As soon as they discovered the random mechanism that underlies the generation of the throws, their mind sought to assimilate this employing of causal explanations (Prodromou, 2008) to operationalize randomness. They began adopting stochastic language when randomness becomes a part of their articulations.

References

Australian Curriculum, Assessment and Reporting Authority. (2010). Australian Curriculum: Mathematics. Version 1.1. Retrieved March 15, 2011, from http://www.acara.edu.au

Glaser, B. G. (1978). Theoretical sensitivity: Advances in the methodology of grounded theory. Mill Valley, CA: Sociology Press.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. International Statistical Review, 65(2), 123-165.

Noss, R., & Hoyles, C. (1996). Windows on mathematical meanings: Learning cultures and computers. London, England: Kluwer Academic Publishers. http://dx.doi.org/10.1007/978-94-009-1696-8

Prodromou, T. (2008). Connecting thinking about distribution (Unpublished doctoral dissertation). University of Warwick, Warwick, UK.

Prodromou, T., & Pratt, D. (2006). The role of causality in the Co-ordination of the two perspectives on distribution within a virtual simulation. Statistics Education Research Journal, 5(2), 69-88.

Robson, C. (1993). Real world research. Oxford, England: Blackwell.

Watson, J. M. (2006). Statistical literacy at school: Growth and goals. , NJ: Lawrence Erlbaum Associates.