MULTIDIRECTIONAL MODELLING FOR FOSTERING STUDENTS' CONNECTIONS BETWEEN REAL CONTEXTS AND DATA, AND PROBABILITY DISTRIBUTIONS

Theodosia Prodromou
University of New England, NSW, Australia
tprodrom@une.edu.au

The article investigates how 14- to 15- year-olds build informal conceptions about data distributions, and theoretical probability distributions as they engaged in a multidirectional modelling process using computer-based simulations. The students of this study are engaged in modelling. First, students examined data from an unknown stochastic process and built a model of the processes that might explain the outputs. Second, the students constructed representations that generated data whose distributions were well predictive of real world samples. This study shows shifts in the conceptual structures across the two directions and points to the potential of specific aspects of multidirectional modelling for fostering the development of students' robust knowledge of the logic of inference when using computer-based simulations to model and investigate connections between real contexts and data, and probability distributions.

BACKGROUND

The mathematical modelling process is a significant activity that can be useful in both developing and applying mathematical content. There is often a need to obtain a "mathematically productive outcome for a problem with genuine real-world motivation" (Galbraith & Stillman, 2006, p. 143). The use of mathematical modelling for solving problems in a real world environment lies "at the heart of functional mathematical literacy" (Burkhardt, 2007, p. 180). The modeling process describes how math problems are modeled and solved (Figure 1).

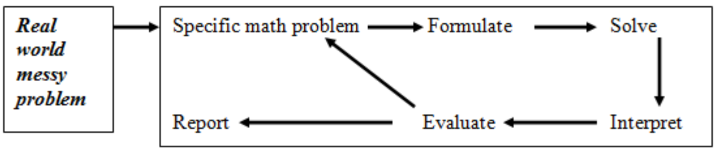

Figure 1: Modelling process (Galbraith, Stillman, & Brown, 2010, p. 135)

This process of modelling an aspect of phenomena in our world reinforces use of models that are formalized in a symbolic system and developed to explain some aspect of our world. Such modelling of some aspect of our world is a process that involves

> a variety of diagrams, concrete models, experienced-based metaphors, and other expressive media-in addition to technical spoken language and symbol systems, each of which emphasize some aspects (but deemphasize others) for the 'thing' that they are used to describe, to explain, or design. Furthermore, model development often involves dimensions of development such as intuition-to-formalization, concrete-to-abstract, situated-to-decontextualized, specific-to-explicit, global-to-analytic. (Lesh & Fennewald, 2010, p. 7)

The modelling process of real world phenomena or situations cannot always be deterministic. We have also probabilistic models that incorporate uncertainty or random error in a formalized way. These probabilistic models, according to their inherent rules, are expected to simulate the behavior of random phenomena and also predict specific properties of random phenomena. For instance, random generators can be seen as determined if only one was aware of a set of factors that causally affect the behavior, in which case the outcomes would be entirely predictable. In practice, this is an unlikely state of affairs and it is likely that one would be interested instead in adopting a probabilistic model.

In this framework 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 were collected from a population, enabling us to compare data from a real observation of this population with a theoretical distribution. Several researchers emphasize the importance of drawing connections between theoretical or real phenomena, models of such phenomena, and data from the

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

real world or generated by the model (e.g., Konold & Kazak, 2008; Lee & Lee, 2009; Lehrer, Kim & Schauble, 2007; Prodromou, 2008; 2012a; 2012b; Stohl & Tarr, 2002). These researchers examined the role of the notion of probability distribution through data modelling as a way to draw connections between data and probability topics in curricula (e.g., Kazak, 2006) and investigate specific aspects of multidirectional modelling when constructing a bidirectional link between the distribution of experimental data and the distribution of the theoretical data (Prodromou, 2008, 2012b).

Working with a variety of digital tools, these researchers have either investigated (a) situations that required students to examine data from "messy" real-world phenomena or derived from an unknown stochastic process, and to build a model of the phenomena/processes that might explain the outputs, or (b) situations that required students to construct models or observe the generation of data from density curves whose distributions were well predictive of real world samples.

These researchers have shown that modelling practices involve making sense of real-world phenomena and defining attributes of the phenomena that can be measured. The development of a model within a digital tool is the most important and ambitious part of the modelling process, because it requires that attributes of the phenomena be defined. However, once such a model is designed, the digital environments afford the collection of data from a model that in turn can be used to test that same model by comparing model-generated data to real-world data. A model can also be used to illustrate aspects of a real-world problem that may not be possible without a data-generating model.

This paper presents data from a study in Australian schools, focusing on how 14- to 15-year-olds develop informal conceptions about data distributions, and theoretical probability distributions as students engaged in a bidirectional reasoning process while using TinkerPlots2 computer-based simulations. At the core of students' reasoning were the modelling processes involved when students observed data from a real-world phenomenon that follow a model and when students constructed a model of the phenomenon that generated this data. More specifically, in this study, students were introduced to theoretical and experimental data of Australian monthly mean temperatures for September 2012 to October 2013, implicitly through a simulation-based approach, and chance variation as an idea was assessed in an intuitive fashion.

THEORETICAL FRAMEWORK

I draw heavily on two lines of research. The first explicates a framework introduced by Prodromou (2012b) to focus the design of the instructional tasks used by students when engaged in collecting, analysing, and interpreting components of this framework in order to relate results from empirical data from a concrete or a computer-simulated two dice rolling experiment with the expected results based on a model of a sample space (based on the analysis of sample space composition), when rolling two dice. This framework that described the process of bidirectional modelling involves the following components: 1. Posing a question, 2. Collecting data, 3. Analyzing data, and 4. Interpreting the results (Friel, O'Conner, & Manner, 2006). Moreover, I draw on research that introduces four primary phases of inferential reasoning when students use computer-based simulations to model and investigate statistical questions (Prodromou, 2013).

To answer the research question about how students develop a bidirectional reasoning about quantities represented in data distributions of sample probability populations, and theoretical probability distributions as students engaged using TinkerPlots2 computer-based simulations, I explicate a framework that will focus on the phases of bidirectional reasoning in the modelling and simulation process when students when engaged in collecting, analysing, and interpreting components of this framework (see figure 2).

In one direction − the "real world data" branch of the framework−, when students begin with examining the distribution of data from an unknown stochastic process or from a real-world problem, they pose questions about particular aspects of data so they will be able to better describe, summarize, compare and generalize data within a context. The second component, collecting data, includes a variety of data collections from populations and samples or generated from a probability experiment. Analyzing data, the third component, involves perceiving a set of data as a distribution when a probability experiment is performed. The fourth component, interpreting the results, encompasses making decisions about the question posed within the context of the problem based on experimental data, and making inferences about a probability distribution. The fifth component, evaluating the model, involves evaluating whether the probability distribution explains the data (that I considered it as a model) and compare the behavior of the model to observed data. If the model fully explains the data, then the decisions about data and characteristics of data are reported. If the model does not fully explain the data, then the modelling process is repeated until a model that fully explains the outputs is found.

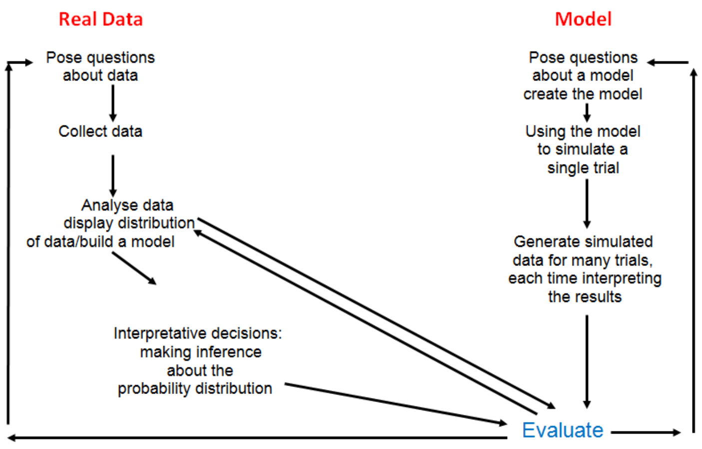

Figure 2: Framework describing the process of bidirectional modelling

In the other direction− the “model” branch of the framework−, when students are engaged in constructing representations that generate data whose distribution are well predictive of real-world samples, they pose questions about a model that will generate data to simulate the real-world problem. This first component involves using a software interface that relies on signal, variation, and spread of data to construct the model. The second component involves using the model to generate a single trial of the simulation, investigating the outcomes from a single trial, constructing an appropriate representation of the outcomes from the single trials. The second component involves analyzing the outcomes from a single trial and considering possible outliers and other individual cases. The third component involves using the model to generate simulated data for many trials, each time interpreting the results of the simulated outcomes while using the observed distribution to assess particular outcomes. The fourth component, evaluating, involves evaluation of the model by comparing its behavior to real data. Such a comparison may prompt them to make interpretative decisions and then either decide to complete the investigative cycle or to continue the investigative cycle by changing the model, and then repeating the phases of the modelling process.

Through several iterations of all phases, in both directions, the students can develop their understanding by beginning to integrate the different elements of their experience into a set of conceptual connections between model and real-world data.

METHODOLOGY

Thirty students in Grade 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 Tinkerplots2 during regular mathematics lessons. All students were familiarised with the TinkerPlots2 software, focusing on its use. In the first session, students watched instructional movies about how to build a simulation in TinkerPlots2. In the second session, students participated in a number of introductory activities related to building a model that simulates real phenomena. They also ran a simulation, observed the generation of data, and explored the output data. Ten students volunteered to spend two extra sessions, outside of class time, to engage in the task reported in this study. In these sessions, students used TinkerPlots2 to construct a simulation of some climatic data for January to December 2013, in the students’ home city.

In the first session, the students used pre-defined probability distributions to draw a density curve that Tinkerplots2 used to generate temperature data. After constructing their model, students were asked to run the simulation and interpret the outcomes. In the second session, students looked for real-world data from the Australian Bureau of meteorology and observed monthly mean maximum and mean minimum temperatures for September 2012 to October 2013.

Each session lasted approximately 45-60 minutes and each pair of students worked directly with the researcher. The researcher interacted continuously with the students in order to observe the reasoning they used to explain the data and simulations. Data collected included audio recordings and video recordings of the screen output on the computer activity. This paper reports work of one pair of students engaged in the modelling process of the “model” branch of the framework.

RESULTS

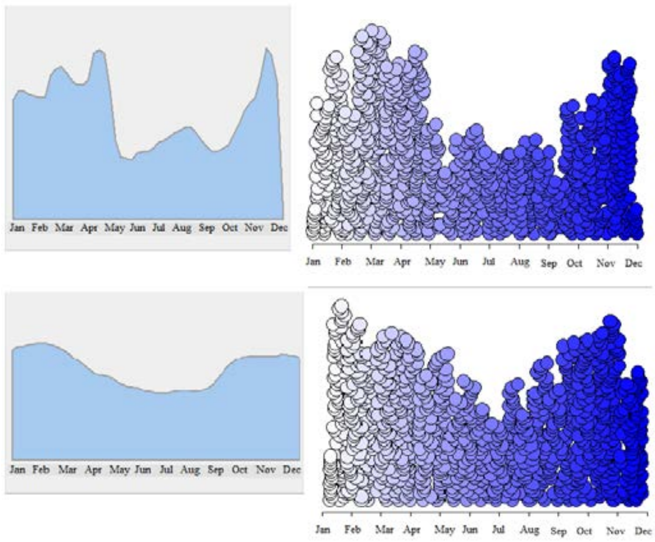

Figure 3. Students’ initial (3a, top left) and redrawn (3c, bottom left) models of mean temperatures and subsequent data generated from each model (3b, top right, and 3d, bottom right).

The session began by having students creating a model (Figure 3a) that generates data that simulate the monthly mean temperature for January to December 2013, in Australia. First the students posed questions about mean monthly temperatures in Australia to gain information on which to build their model. They suggested grouping the months into seasons. They stated that during December, January, February and March the mean temperature is high; in April the temperature drops; in May and June the temperature drops dramatically; in July and August, that temperature drops very low; and that during September, October, and November the temperatures rise again. One of the students in the pair wanted to include very high and very low temperatures that would change the model. The other student added that temperature anomalies are included in the mean temperature. After their discussion about the signal, variation, and spread of data involved in the model, the students engaged with using a software interface to construct the model (Figure 3a). After generating mean temperatures for January 2013 to December, the students looked at the distribution of the monthly temperatures (Figure 3b) and they began interrogating the context of their personal experiences. They used TinkerPlots2 to generate simulated temperatures from the model, considering each time possible outliers/temperature anomalies and other individual data cases. During the processes of interpreting the results of the simulation model, while using the observed distribution to assess the outcomes, the students realized that the mean temperature for December in the generated data was lower than its real counterpart, and the mean temperature for April was higher than the real temperatures they personally experienced. The model (the density curve drawn in the sampler) appeared to need changes of mean temperature for the months December and April. This fourth component of evaluating the model required the students to make interpretative decisions in regards to changing the model by redrawing the density function of the monthly mean temperature in Australia. This redrawn model (Figure 3c) was used through each phase of the framework and produced computer-generated data that the students judged as representative of the real data. Nonetheless, the pair was not ready to make a final decision about the adequacy of the model and complete the investigative cycle.

Students articulated that they were not able to consider the goodness of the model, if they would not be provided with opportunities to observe the actual monthly mean temperature for January 2013 to Dec 2013 in the original distribution of data and the model-generated data. When they attempted to find information about weather during the 14 months from September 2012 to October 2013, from the Australian Government Bureau of meteorology website, they learned that October temperatures were unusually warm compared to the 1961-1990 mean. The students looked closely at the temperatures in October of previous years. When they compared their model to the graphical representations of the temperatures in October of 2011 and October 2010, they concluded that the model they built was okay because it matches common temperature variation, but not the particularly warm October 2012. Students focused their attention on examining monthly temperatures of previous years that broke the national record for the warmest 12 month period because they believed that if they were able to detect a temperatures pattern in the Australian's climate change over the last years, they could make predictions about the future temperatures.

The one student of the pair argued that since this year the mean spring temperatures broke the national record for the warmest 12-months period, it was impossible to create a model that would accurately simulates annual temperatures for the next years. He added that if the temperatures keep on increasing over time they do not know how many degrees Celsius would be increasing.

The other student looked at the global mean temperature over the last 100 years and added that the global mean global temperature has increased by around 0.74 °C, and this rate of warming was very unusual in the context of natural climate variability. Perplexed about the vagaries of the weather, the students decided to increase the Spring and Summer temperatures in in their model. However, the one of the students explained that although the record for Australia's warmest calendar year is currently held by 2005, there is no guarantee that 2013 will go on to be Australia's hottest calendar year on record, because as the temperatures decreased after 2005, the temperatures for the remainder of the year may be decreased.

DISCUSSION

The students at the first phase of the modelling process relied explicitly on their personal experiences to draw a density curve that Tinkerplots2 used to generate temperature data. They recalled common temperatures for several months and implemented in their model common temperature variation. The design of the model requires a fairly good co-ordination of common temperatures of several months (signal) and common temperature variation (noise), so that modelers can make "optimal" choices when designing a model that will generate sample data that resembles as much as possible the actual monthly temperatures for January 2013 to December 2013. The way students expressed the relationship between common temperatures for several months (signal) and common temperature variation (noise) is of vital importance while they are working with real temperatures and observing climate change and variability of temperatures in Australia. In fact, the reasoning the students do between actual temperatures in Australia and the model that they built for the monthly mean temperature for January 2013 to December 2013 is associated with what they experience in the behavior of the climate change over time.

They sought for information about mean temperature anomalies and graphical representations of Australian monthly mean temperature anomalies for the last months. These temperature anomalies were compared to climate anomalies and national records for the warmest 12-month periods of previous years. Expressions focusing on these comparisons of temperatures from previous different years and the relation between common temperatures and common temperature variation, involve elements of abstraction and simplification of reality with respect to the real situation studied. In this article, the model that students built for the monthly mean temperature for January 2013 to December 2013, emphasizes the mathematization of real situations in meaningful ways for the learner. Although no models that are represented in a symbolic system suitable for probability calculus have been developed and used by students in this study, the modelling activities enable students to draw connections between science and mathematics.

It is interesting what we can learn from the way the students compared their model with the real Australian temperatures. Students' understanding of the connection of science and mathematics is increased by the modelling activities. Students realized that science (climate change) can be used to inform them about the past and to develop mathematical models that make prediction about the future. A computer simulation-based approach to bidirectional modelling, scaffolds the construction of a bidirectional reasoning between theoretical simulations that models and investigates connections between a real context, real monthly temperatures, common temperatures for several months (signal), common temperature variation (noise) and probability distribution that is specified by a density function. These are areas where sophisticated understanding and more expert knowledge in probability distributions, statistical knowledge and temperature trends can be useful to students in decision making and modelling monthly mean temperature for a certain period of time.

REFERENCES

Burkhardt, H. (2007). Functional mathematics and teaching modeling. In C. Haines, P. Galibraith, W. Blum, and S. Khan (Eds.), *Mathematical modeling: Education, engineering and Economics* (pp. 177-186). Chichester, UK: Horwood.

Friel, S. N., O' Connor, W., & Mamer, J. D. (2006). More than "meanmedianmode" and a bar graph: What's needed to have a statistical conversation? In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (pp. 117-137). Reston, VA: National Council of Teachers of Mathematics.

Galbraith, P., & Stillman, G. (2006). A framework for identifying student blockages during transitions in the modeling process. *ZDM*, *38*(2), 143-162.

Galbraith, P., Stillman, G., & Brown, J. (2010). Turning ideas into modeling problems. In R. Lesh, P.L. Galbraith, C. R. Haines & A. Hurford (Eds.), *Modeling students' mathematical modeling competencies* , (pp. 133-1443). New York: Springer.

Kazak, S. (2006). *Investigating elementary school students' reasoning about distributions in various chance events*. Unpublished dissertation, Washington University, St. Louis, MO.

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovation in Statistics Education*, *2*(1), [Online http://www.escholarship.org/uc/uclasta_cts_tise].

Lee, H. S., & Lee, J. T. (2009). Reasoning about probabilistic phenomena: Lessons learned and applied in software design. *Technology Innovations in Statistics* Education *3*(2), [Online http://www.escholarship.org/uclastat_cts_tise].

Lehrer, R., Kim, M., & Schauble, L., (2007). Supporting the development of concepts of statistics by engaging students in modeling and measuring variability. *International Journal of Computers for Mathematics Learning*, *12*, 195-216.

Lesh, R., & Fennewald, T. (2010). Introduction to part I modeling: What is it? Why do it? In R. Lesh et al. (Eds.), *Modeling students' mathematical modeling Competencies*, (pp. 5-10). New York: Springer.

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

Prodromou, T. (2012a). Students' construction of meanings about the co-ordination of the two epistemological perspectives on distribution. *International Journal of Statistics and Probability*, *1*(2), 283-300. doi: 10.5539/ijsp.v1n2p283

Prodromou, T. (2012b). Connecting experimental probability and theoretical probability. *ZDM – The International Journal on Mathematics Education*, *44*(7), 855-868.

Prodromou, T. (2013). Informal inferential reasoning using a modelling approach within a computer-based simulation. *World Academy of Science, Engineering and Technology*, *78*, 1761-1766. URL: http://www.waset.org/journals/waset/v78/v78-301.pdf

Stohl, H., & Tarr, J. E. (2002). Developing notions of inference with probability simulation tools. *Journal of Mathematical Behavior*, *21*(3), 319-337.

TinkerPlots: Dynamic data exploration (Version 2.0) [Computer software]. Emeryville: CA: Key Curriculum Press.