Repurpose and extend: making a model statistical

Michal Dvir, Dani Ben-Zvi

Repurpose and extend: making a model statistical

Michal Dvir and Dani Ben-Zvi

University of Haifa, Faculty of Education, Haifa, Israel; dvirmich@gmail.com

The goal of this article is to examine how learners' repurposing of a model they had previously constructed, possibly by extending its features, can support young learner's reasoning with key statistical features. This process is facilitated by a learning trajectory integrating multiple real world and probabilistic modeling tasks. We introduce our framework describing young learner's reasoning with informal statistical models and modeling (RISM) and focus on a key element in the framework: young learners' expectations regarding the data that can be collected (or generated) from a conjectured population. We offer an illustrative case study of two-sixth graders' reasoning, focusing on a specific model they had constructed and repurposed throughout their investigation, how its features developed and how it ultimately matured to include an underlying probabilistic mechanism. We discuss how this relates to our notion of expected data and close with suggested implications.

Keywords: Statistical model, statistical modeling, integrated modeling approach.

Introduction

There has been a growing interest in the statistics education community in statistical modeling as a pedagogical paradigm to developing learners' statistical reasoning. However, a deeper understanding of what statistical modeling entails is needed to better exhaust its pedagogical potential, especially with regard to young learners. In particular, a better understanding of how novices can be encouraged to construct models that are accompanied with uniquely statistical features, such as accounting for possible variability – that of the data explored, as well as of the sampling procedure that yielded it – is still warranted. The purpose of this article is to explore this, based on an illustrative case study of a pair of sixth grade students' (12 year olds) participation in a uniquely designed learning environment. We track a specific model they had constructed, and gradually repurposed as they participated in various modeling tasks. The new roles assigned to the model were also accompanied by extending and refining some of the models' features, gradually becoming more 'statistical' – that is, it eventually featured uniquely statistical features that distinguish it from a non-statistical model.

Background – The RISM framework and the IMA

A model – a representation serving an explanatory or descriptive purpose (Hesse, 1962) – would serve as a statistical model if it was constructed for a statistical purpose, typically meaning: 1) the phenomenon it is intended to explain has a variability aspect, and 2) utilizing it includes employing probabilistic considerations (Brown & Kass, 2009). The latter is associated with the notion that the description a statistical model offers should be non-deterministic in its nature (Budgett & Pfannkuch, 2015). However, when young learners are the modelers, an informal alternative is necessary, void of the conventional procedures and calculations. Within the context of a statistical inference, a statistical model (formal or informal) needs to serve a dual representative purpose: simultaneously depicting both a conjecture about the explored phenomenon (Konold & Kazak,

2008) and its underlying probabilistic mechanism (Pfannkuch & Ziedins, 2014), as well as the data itself. It is this dual representative purpose that has inspired our framework for depicting young learners' *reasoning with informal statistical models and modeling* (RISM).

The RISM framework suggests describing the statistical modeling process as a continuous development of a set of key elements portrayed in Figure 1. The dual representative purpose of the model is translated into two planes: the data and the conjecture plane. In both, the novice modeler gradually constructs a simplified version of either the explored phenomenon (the phenomenon plane, bottom of Figure 1) or his abstract conjecture (the conjecture plane, top part). The latter may be trivial for an experienced statistician, already familiar with a vast variety of readymade statistical models and formal fit assessment procedures, but challenging for a novice. Alternatively, co-constructing two concurrent models – a data model depicting observed patterns in the data, as well as a conjecture model depicting the data expected to be collected – and assessing their compatibility, can serve as an informal alternative to the practice of model fit evaluation (Dvir & Ben-Zvi, 2018).

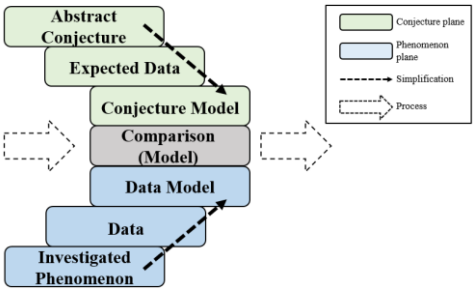

Figure 1: The RISM framework's snapshot

Worth noting is the role the expected data serve in this depiction of the informal statistical modeling process: as the expert statistician's readymade abstract conjecture is a statistical model, it should be already accompanied with previously known considerations regarding the data it can generate, and how the resulting generated data may vary dependent on the type of the model and sample size. In cases where the model is lesser known, the expert would seek to investigate how the data generated from it would behave. However, a young modeler would not typically be aware of the formally expected behavior of such generated data, nor that this behavior may vary and should be explored. Thus by 'expected data' the framework merely refers to the learners' (typically naïve) perception of the data one might collect from the investigated population if his conjecture were true. We believe this expectation is critical when considering whether the investigated conjecture model can be considered a statistical model. For it to be one – the expectation should express some (albeit informal) understanding of the probabilistic considerations associated with generating data from a model.

For this reason, integrating an investigation of the conjecture model's probabilistic mechanism may be beneficial in supporting young learners' reasoning with statistical models and modeling. Therefore, *the integrated modeling approach* (IMA, Manor & Ben-Zvi, 2017) suggests following a 'real world' inquiry with a 'probability world' inquiry, in which the probabilistic considerations of interest becomes the main subject of investigation. The probabilistic inquiry is typically initiated by

students' creating a dynamic model using the TinkerPlots Sampler[1] (Konold & Miller, 2015), based on their real world conjecture models. They then draw multiple simulated (same sized) samples from it, compare between them and eventually construct a sampling distribution of a chosen statistic. The data examined in the probability world is therefore no longer a single sample, rather multiple samples, and the conjecture and inferences made describe the samples' behavior. These inferences, although different in their nature, are closely connected to the real world inquiry that instigated the probabilistic exploration and therefore can be then related back to inform a progression in the real world inquiry.

The real world conjecture model is therefore utilized in the probability world to serve a different purpose (probabilistic rather than real), however what is examined is its underlying probabilistic mechanism, mirrored by the sampling variability students observe. This can lead to refinements of the Sampler model from which the samples were generated, such as adding new attributes to better represent sampling variability (Manor & Ben-Zvi, 2017). Similar extension have also been reported in other integrated modeling endeavors (Lehrer, 2017), however how these occur and how can these be then utilized to inform student's real world instigating inquiries has not been thoroughly examined. This may be in part due to the different visualization students employ when constructing the Sampler model, thus it is unclear how its extensions may manifest in its original real world representation. Therefore, we have chosen to focus on a specific representation first utilized in the preceding real world investigation but also employed during the subsequent probabilistic inquiry, and examine: *How can young learners repurpose a real world conjecture model and extend some of its features in a follow-up probabilistic inquiry?*

Method

This study was conducted as part of the Connections project, a longitudinal design and research project (began at 2005) aiming at promoting young learners' statistical reasoning, in a technologically-enhanced and inquiry-based learning environment. We focus on the data collected in 2016 of a pair of second year participating 12 year-old students from a public school in Israel. Their entire sixth grade class participated in a 19 lesson long learning trajectory, which included three full (data and probability) investigation cycles. The students used TinkerPlots (Konold & Miller, 2015) to both create their data representations, as well as design a generative conjecture model and draw multiple samples from it. The findings we present in this article were taken from the second investigation cycle in which students were asked to make inferences about 1300 students of three schools in the district (including their school) based on a random sample of 60 students. Since our purpose was to learn more about the students' reasoning process employed to elucidate the role of a specific phenomenon, we chose to focus on an illustrative case study. We focus on the reasoning employed by a pair of academically successful students, Noa and Ido, second year participants in Connections. They were selected as they were verbal and communicated their reasoning openly and fluently, and exhibited a behavior observed in several other pairs while disclosing much about the reasoning that encouraged it: they utilized the same visualization in their

[1] The TinkerPlots Sampler allows students to design and run probability simulations, subsequently plotting the results to give a visual representation of the outcome over many samples (a sampling distribution).

inquiry of the real and probabilistic worlds. This facilitated us to more accurately examine how the same model was repurposed and what model features were subsequently extended.

All the activities and participants' actions were videotaped by three cameras during class discussion and Camtasia software to concurrently document the students' computer screen and articulations during their pair work. After transcribing all the pair's videos, the resulting data corpus was reviewed and analyzed according to the interpretative microgenetic method (Siegler, 2006), utilizing the RISM framework to translate the convoluted statistical modeling processes observed to discrete sets of well-organized snapshots (Dvir & Ben-Zvi, 2018). Key scenes and their transcripts, were discussed and triangulated (Schoenfeld, 2007) amongst the co-authors and a third Connections researcher.

Findings

This section illustrates the progression of a specific model the pair had constructed during the second investigation cycle. Figure 2 illustrates the progression we will describe, from its first appearances during the real world investigation, to its development during the following probabilistic inquiry.

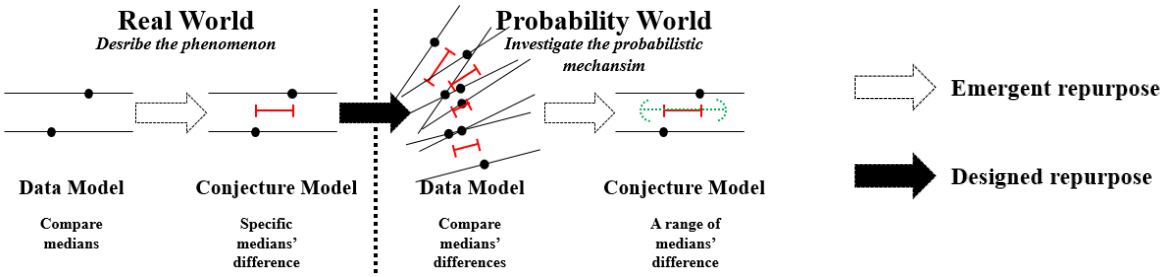

Figure 2: Illustration of repurposes and extensions of the medians model

Initial use of the medians model: repurposing a real world data model to a conjecture model

The second investigation cycle focused on "sportiveness and gender". Focusing on students' 600 m running times, Noa formulated their research question: *Is the stigma that girls run slower than boys true?* The pair formulated an initial conjecture: "there is a small difference between boys [and girls, in favour of] the boys", and began to investigate their real data by examining the 600 m results by gender (Figure 3). Asked by the researcher what can they see in the data, Noa replied:

Noa: That we [girls] do not run so well… 2 minutes and 5 seconds [miscalculation of the boys' median – 165 sec] is not so good… [and the girls] run even worse. 2 minutes and 13 seconds [miscalculation of the girls' median – 173 sec].

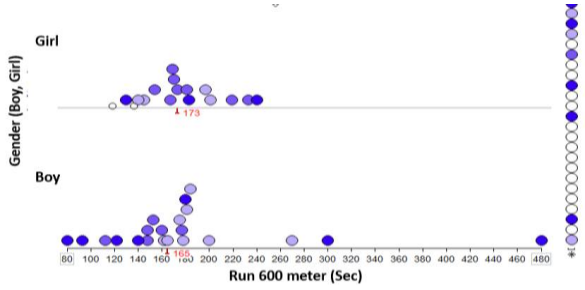

Figure 3: 600 m running times (seconds) by gender, colored according to grade

In response to the researcher' prompt, Noa described the median of each gender, and compared between them stating "[the girls] run even worse." This utilization of the medians, here representing what Noa *saw* in the data (her data model) is what we refer to as '*the medians model*', and its progression is what we will focus on throughout the Findings Section. The researcher, in accordance with the learning trajectory design, encouraged the pair to infer beyond the data, leading Noa to revisit her conjecture:

Noa: I think it [the population of 1300 students] will be less extreme than what I said. There would be a lot more boys than girls that will enlarge the median…

R: And who will be better, boys or girls?

Noa: Boys… I think that the boys' median will be a bit bigger and the girls median will be a little more bigger [means: smaller] … by 1 [second]

The underlying behaviour of Noa's current conjecture (her abstract conjecture) had not strayed much from her initial conjecture ("there is a small difference between boys [and girls, in favour of] the boys"). The novelty, however, is the model *repurposed* to describe it, as for the first time the medians model is awarded this role. Furthermore, Noa also described here not only which of the medians will be "bigger", but also by how much – a new extension to the medians model. This implies that the data Noa expected her conjectured population to generate is rather specific. The extended feature of acknowledging the actual size (although not necessarily by number) of the difference between the two medians, will gradually take on a bigger role, reflecting a development in the pair's perception of the expected data.

Repurposing in the probability world

During the remainder of the lesson, as well as the one that followed, the pair moved on to the next task of designing a generative model using the Sampler. Their model was based on both the real data sample as well as their perception of the running results of the population of 1300 students, thus reflecting their current conjecture: the girls' and boy results are almost equal, but boys run a little bit faster (Figure 4). After constructing the Sampler model, the pair began to generate simulated random samples from it. In the first sample drawn, the girls' median was significantly smaller than the boys' (Figure 5). Noa addressed this in multiple ways, for example, when the researcher introduced the boxplot in the following lesson. As the researcher asked the pair how the boxplot could be utilized, Noa replied: "We can see where there are more data, and also who is better, because if, for example our [the girls'] median is here [relatively small], and their median [the boys'] is here [larger] it means that we [the girls] are better here [in this sample]". Noa's response indicated that although the addition of the boxplot was an extended feature of the visualization, the underlying data model she was still utilizing is the medians model. Although now analysing a simulated sample, the role this model seemed to serve is similar to that in prior utilizations of the medians model: describe the data (a single sample) at hand. After drawing a second simulated sample, the researcher asked how the pair would like to compare between the samples.

Noa: Here too [in the second simulated sample as well as the first], we [the girls] are better [than the boys]

R: Because of what?

Noa: Because again [in another sample we see that] the [girls'] median is at better results and

Ido: Why? But our [the boys'] median is becoming closer to yours [the girls' median]

Noa: It is, but still [the girls' median is better than the boys']…

Ido: It [the difference between the medians] is very similar, all the time…

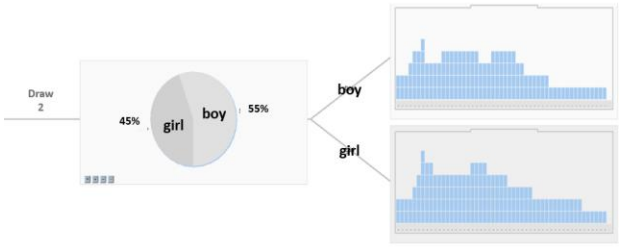

Figure 4: The Sampler model – the attribute on the left is gender and on the right, two "almost equal" distributions were designed for the boys' (top) and girls' (bottom) 600 m running times

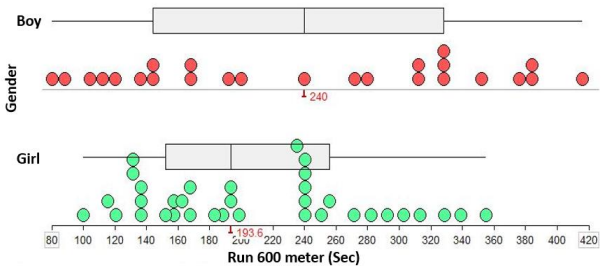

Figure 5: Adding a boxplot to investigate the first simulated sample of the boys' (top) and girls' (bottom) running times in which the girls' median is smaller than the boys'

Noa began by stating the result of the comparison: "here too". As the researcher prompted her to explain how she had ascertained this result, Noa referred again to the median as an indicator regarding who runs better. However, implied in her response is a possible extended feature: while comparing the medians (which is smaller – the boys' or the girls') served as a model for describing what the data in a single sample indicated (regarding who runs better), comparing which of the two medians is "better" across the two samples was a method to compare between the new simulated sample and the previous one. This *repurposing* was implied both by her use of "again" as well as the researchers' question that instigated this exchange (how would they like to compare between samples).

Ido's reaction introduced a new aspect, more explicitly addressing the new type of comparison (between samples, rather than between the two distributions within a single sample): comparing differences. Not only was he acknowledging the size of the difference between the two medians, as opposed to Noa's acknowledging only which of the two prevailed, he was also explicitly focusing on the change that accrued between this current sample and the former: the boys' median was "becoming closer" to the girls', meaning the difference is now smaller. Despite Noa's reply showing she did not consider the smaller gap between the two medians as a relevant indicator, Ido's final statement takes this new aspect even one step further, as he used a much more general formulation: "all the time". However, as the pair's inquiry progressed the extended feature Ido had introduced was indeed picked up by Noa. An example for this can be seen in a later exchange:

R: What will happen if we draw another sample?

Noa: It seems to me it [the new sample] is the same. The boys [their median] are slowly slowly getting closer to us [to the girls' median]

This last statement shows that not only was Noa now considering comparing the differences between the medians as a means of comparing the new sample with the previous samples, it is now even becoming a part of her newly forming conjecture. However, this conjecture does not reflect Noa's view of the real world phenomenon the pair were investigating (who runs better), rather a different kind of phenomenon: how much do samples such as those they were investigating vary? We therefore consider this to be a *repurposing* of the original medians model, here also serving as a conjecture model for the probabilistic phenomenon the pair were investigating.

The pair continued to generate more simulated random samples, and for each compared the difference between the boys' and girls' medians with the differences they saw in other simulated samples: "you [the boys' median] are getting closer to us [the girls' median] but it is taking you a lot of time" (Noa), "now it [the difference between two medians] is almost equal [to the previous difference]!" (Ido). The changes they observed varied, to the extent that in one sample, much to Noa's disappointment, the boys' median even surpassed the girls'. After several samples were drawn, the researcher prompted the pair to consider whether a sample size 60 can be trusted:

Noa: No, because let's say once they [the boys' median] can be here [at one point] and we [the girls' median] can be here [another point], and another time they [the boys' median] can be here [different point] and we [the girls' median] can be here [another different point]. It's confusing.

Ido: It [the difference between the medians] can always change.

Noa: Here [in other samples] it is like that they [the boys' median] slowly slowly closed [the gap] to us [the girls' median], and here [in one sample] they [the boys' median] are before us [is smaller than the girls' median].

Asked to evaluate the trustworthiness of samples size 60, both students responded by referring to the changes they had observed across the simulated samples: While Noa initially referred to variations in which of two medians prevailed, it seems that Ido was still focused on the difference between the two, reflected in his use of the singular form 'it' rather than the more appropriate 'they' had he been referring to the two medians. Nevertheless, Noa's final statement showed that she too was considering changes to the differences ("slowly slowly closed [the gap] to us"). Noa's depiction describes a range of possible differences between the two medians, a long way from her original expectation of one specific result (bigger "by 1"). The range of possible results is a key extended feature of the medians model, making it now much more statistical than its initial version.

Discussion

Focusing on a specific model that emerged during the pairs' real world inquiry – the medians model – we sought to explore how the pair *repurposed* its goal and *extended* its features as they progressed in their inquiry and transitioned into the probabilistic follow up inquiry (Figure 2). As Noa and Ido designed their Sampler generative model to reflect their initial real world *abstract conjecture* (regarding running results), drawing multiple random samples from it allowed them to gradually explore its underlying probabilistic mechanism. Because they chose to repurpose the same visual representation (the medians model) rather than create a new one, they were both learning how the real world expected data would behave (or vary), and how this behavior would be portrayed in their

model (what can happen to the two medians). This indeed required extending some features of the original medians model, possibly as its original use (check each median and compare between them) was too complicated to track over several samples. The end result of this process was a much more elaborate and statistical model – one that serves the dual representative purpose in the real world investigation (both data and conjecture), and also is accompanied by a non-deterministic expectation regarding the data it can generate. Although the pairs' choice to repurpose a model they had used in their real world inquiry might appear coincidental or idiosyncratic, we have observed several other pairs chose to do the same. We attribute this to the IMA design: initiating the learning trajectory with a real world inquiry, and its close connection to the probabilistic follow up investigation. This case study illustrates how integrating a follow-up inquiry regarding probabilistic concerns that emerge from a real world investigation within an informal inferential setting can promote an important aspect of young learners' RISM: experiencing the population as a generative model, one from which various samples can be generated, and cultivating the learners' understanding that this variability needs to be examined and accounted for. In particular, repurposing the same models throughout both types of inquiry, can support reasoning with its underlying probabilistic mechanism.

We thank the *Connections* team, the University of Haifa, the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation Grant 1716/12.

References

Brown, E. N., & Kass, R. E. (2009). What is statistics? *The American Statistician, 63*(2), 105–123.

Budgett, S., & Pfannkuch, M. (2015). Building conditional probability concepts through reasoning from an eikosogram model: A pilot study. In *Proceedings of SRTL9* (pp. 10–23). Paderborn, Germany: University of Paderborn.

Dvir, M., & Ben-Zvi, D. (2018). The role of model comparison in young learners' reasoning with statistical models and modeling. *ZDM - International Journal on Mathematics Education*. doi:10.1007/s11858-018-0987-4.

Hesse, M. B. (1962). *Forces and fields: The concept of action at a distance in the history of physics*. Mineola, NY: Dover.

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education, 2*(1), Article 1.

Konold, C., & Miller, C. (2015). *TinkerPlots* (Version 2.3.1) [Software]. http://www.tinkerplots.com/

Lehrer, R. (2017). Modeling Signal-Noise Processes Supports Student Construction of a Hierarchical Image of Sample. *Statistics Education Research Journal*, *16*(2), 64–85.

Manor, H., & Ben-Zvi, D. (2017). Students' emergent articulations of statistical models and modeling in making informal statistical inferences. *Statistics Education Research Journal*, *16*(2), 116–143.

Pfannkuch, M., & Ziedins, I. (2014). A modelling perspective on probability. In E.J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: Presenting plural perspectives* (pp. 101–116). Dordrecht: Springer.

Schoenfeld, A. H. (2007). Method. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 69–107). Charlotte, NC: Information Age Publishing.

Siegler, R. S. (2006). Microgenetic analyses of learning. In D. Kuhn & R.S. Siegler (Eds.), Handbook of child psychology: Cognition, perception, and language (Vol. 2, 6th ed., pp. 464–510). Hoboken, NJ: Wiley.