CONSTRUCTING SIMULATIONS TO EXPRESS DEVELOPING STATISTICAL KNOWLEDGE

Lulu Healy
PUC, São Paulo, Brazil
lulu@pucsp.br

This paper reports on an attempt to involve mathematics teachers, with a limited previous experience in exploring statistical concepts, in the collaborative design of computational tools that can be used for simulating data sets. It explores the constructionist conjecture that the design of such tools will encourage designers as learners *to reflect upon the statistical concepts incorporated in the tools under development, since generating data-sets on the basis of different characteristics, such as average, spread, or skewness, necessitates the making explicit of thinking related to these notions and the construction of some sense of random processes. It describes how involvement in the design process involved participants in coming to see distributions as statistical entities, with aggregate properties that indicate how their data is centred and spread.*

INTRODUCTION

Statistics has only recently entered the Brazilian school mathematics curriculum in any substantial form. Data handling, encompassing statistics, probability and permutations, now represents one of the four "topic blocks" prescribed in the National Curriculum guide (PCN, published in 1998). In relation to statistics, the curriculum guidelines stress the importance of involving learners in the collection, organisation and communication of data, using tables, graphs and recommend the introduction of statistical measures such as the mean, median and mode as elements to be employed in the interpretation of statistical data. Despite these curriculum demands, statistics education has not featured highly (if at all) in teacher education courses, meaning that reasoning statistically remains an area with which Brazilian mathematics teachers are still very unfamiliar and insecure and there is every indication that they are not sufficiently prepared to create the kind of classroom environments being asked of them.

Although research into data handling competencies within the Brazilian school context is relatively rare, a feature of the studies now beginning to emerge is a focus on the use of computational tools as a means to enable approaches to statistics that emphasise the exploration, analysis and interpretation of data sets and their properties (Santos, 2003; Costa, 2004). Following this trend, one option for teacher education is to create technology-integrated learning activities, which model the kinds of activities we would like teachers to use in their own classroom. A problem with this strategy is that the teachers bypass the design phase, with the result that they may not feel ready to come up with new activities of their own or even to adapt existing activities according to the particular needs of their students. An alternative strategy, which we are currently investigating within the research group *Technology and Media for Mathematical Expression* (TecMEM) of PUC São Paulo, is to involve mathematics teachers, together with researchers and computer programmers, not only in the design of activities to encourage statistical reasoning, but also in the design of the computational environments which form the context in which this reasoning is to take place. The constructionist idea on which this approach is based is not that we create well polished "finished" software, rather that we create microworlds that represent our tinkering and can be subsequently tinkered with by others (Papert, 1991).

This paper presents a brief synopsis of the strategies that emerged as we attempted to develop, collaboratively, one such environment, a computational microworld in which ideas related to average and spread can be explored and expressed.

THE STORY BEGINS…

For most of the teachers involved in TecMEM, the most familiar statistical measure (even the only familiar measure for some) is that of the arithmetic mean. This seemed like as good a starting point as any. But, as Stella (2003) reports, the dominant view in Brazilian mathematics classrooms is that of mean as algorithm, its meaning synonymous with the mathematical operations used for its computation. To counter this view, rather of presenting students with the

problem of calculating a mean from a given set of data, Stella experimented with some of the construction problems described in Mokros and Russell (1995) – problems in which instead of calculating the mean for a given data set, given the size of the data set and its mean, students have to suggest possible distributions. One of the members of our research group was particularly taken with this kind of problem, but she argued, in the paper and pencil context, it is pretty fiddly to arrive at possible data sets, making it unlikely that students would suggest more than one distribution to fit the given constraints. She brought this problem to group members, asking "*couldn't we use the computer to calculate possible data sets*"? This seemed like an interesting challenge, and a challenge that suggested two questions for research:

- Would participating in design of tools for simulating data sets encourage designers *as learners* to reflect upon statistical concepts incorporated in the tools under development?
- Would the design process encourage designers *as teachers* to reflect upon the kinds of representations that might permit their students to access and explore these same ideas?

The work related to this challenge is still ongoing and the microworld far from finished, the remainder of this paper concentrates primarily on the first of these two questions, as the strategies that emerged in two of the groups who worked upon this challenge are described. Since our starting point was an aggregate feature of the data set, it seemed reasonable to conjecture that the design of tools for simulating possible distributions which have this feature would encourage a view of a data set as mathematical entity in its own right (as opposed to the common perception of a data set as a collection of individual data set described, for example, by Hancock, Kaput, and Goldsmith, 1992).

Because of the emphasis on design as learning, we decided not to work with existing statistical software tools, but to program our own, using the software *Imagine Logo*. [A Portuguese version of the software was used, but for this paper the microworld and code has been translated into English.] Not all of the members of our group were familiar with this software, however among the eleven TecMEM members who expressed an interest in this project, we could count on four members with programming experience, whose role was to coordinate the formalization of the ideas of the rest of the group. Before we started work on the given challenge, the Imagine tool `random` was introduced to all eleven participants. This tool, given a positive whole number as input, outputs an integer between 0 and one less than the number (i.e., the output from `random 5` is is `0`, `1`, `2`, `3` or `4`).

REFINING THE CHALLENGE

Having accepting the challenge, we split into four smaller groups to work on solution strategies, but almost immediately we had to reconvene as it became clear that "the rules" were not entirely clear. Though some were happy to think about the problem completely in abstract terms, for others it was important to ground the challenge in a particular situation. We returned to the Mokros and Russell (1995) paper, and the construction problem based on a set of 8 families with a mean size of 4. This problem situation became the reference context for much of the subsequent discussions. Because of this and because of the way the Imagine tool `random` works, it was decided to limit the simulations to data sets involving whole numbers.

It also became clear that in order to build a computer model of the situation, it would be necessary to specify not only the mean value of the data set to be generated, but also its spread, or at least the minimum and maximum possible values. At first, some of the group were worried that this new demand was not part of the original challenge, but rather a constraint imposed by the 'computer.' Daniel, one of the programmers of the group, however pointed out that even in the paper-and-pencil version of the task, minimum and maximum values were necessarily chosen and what was different in the computational context was that these properties of the data set were afforded a rather more explicit role in the task of simulating data sets. His argument was convincing and other group members began to agree that having these as 'up front' variables that could be fixed (if we wished to examine different data sets with the same mean and spread) or altered (when interest was on data sets with the same mean but different maximum and minimum

values) might help learners – and help them – think about the relationships between spread, mean and the shape of a distribution.

We also discussed some interface issues and decided that what we wanted to produce on screen was a list containing the values of the generated data set as well as a frequency plot. As the rest of the group split into four and worked on strategies for generating the data-sets, I concentrated on producing a first version of such a graphing tool. Each of the four groups, as well as being assigned a programmer, also contained one participant who was ascribed the role of recorder. Their task was to act as participant-observer in the group, noting the strategies that developed, as well as how they were tested and modified.

A PROBLEM WITH RANDOM

One of the groups began by working with a strategy which involved using the `random` tool to pick possible values for all but the last value in the data set. The chosen values could then be summed and a final value added to make the total equal to the given mean multiplied by the given sample size. This was expressed in the group's report as follows:

> *As we have eight families, if we choose at random a number between a minimum of one and a maximum of eight for the number of people in each of the first seven families, then we can calculate how many to put in the eighth family. With a mean of four, it has to make the total number of people thirty-two.*

The sub-group's programmer, Leila, had the task of expressing this as a *Logo* procedure; in the meantime, Edith and Marcia came up with several possible lists of their own. Perhaps it is worth noting that the main proponent of this strategy was the programmer and in constructing possible lists with paper and pencil, the others did not follow this strategy, but rather used a strategy based on the property that the total sum of deviations from the mean is zero (they developed an iterative process which involved choosing a value, considering its difference from the mean and then selecting several other values which taken together compensate this difference). This strategy was also used and described by other group members and after the initial data generation procedures were written, Daniel and Carlos, two of the programmers, set themselves the task of modelling this strategy. This has turned out to be a considerable programming challenge, which had not been completely solved at the time this paper was written. As the pair chose these values, they also discussed the particular situation of family size, relating their choices to families they knew and to their belief that, on the whole, large families are becoming rarer in much of Brazil. They hence chose data sets in which the values were in general clustered close to the mean.

The group reconvened when Leila's procedure had been completed. The first data set produced by Leila's procedure was [4 1 8 8 4 1 5 1], Leila was pleased and suggested they up *n* (the number of families), they tried 40 and again the procedure output a 'legal' data set (Figure 1).

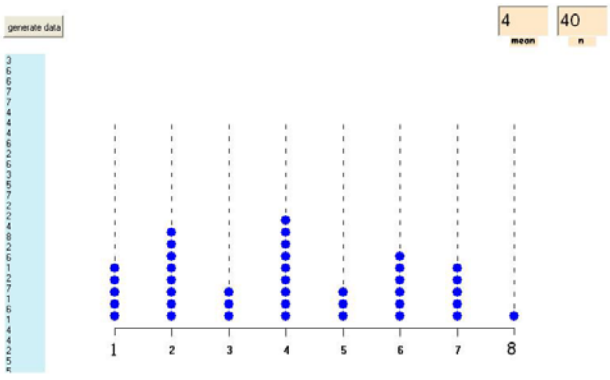

Figure 1:A data set representing the size of 40 families, for which the mean size is four

When they increased the number of families to 100, however, the last number in the data set was -125. Not only well out of the permitted range, but also completely impossible given the context of

family size! They returned to test the procedure with smaller numbers of families and it quickly became clear that in the majority of cases the last number was a negative value. It was not hard for the group to see that one problem with their method was that the total of n–1 families tended to be too far out for the last value to 'correct' the total of the completed data set using a value within the range of permitted values that they had decided upon – but what particularly intrigued the group was why the last value tended to be a negative one.

Playing with the maximum value and then the value of the mean helped them make sense of why this happened. First, they discovered that decreasing the maximum value by one (from eight to seven) increased the chance of obtaining a 'legal' data set, while larger changes in the same direction had the effect the last value would be out of the permitted range but positive rather than negative. They hypothesised that the incidence of 'legal' data sets would also be increased if the mean value was the midpoint between the maximum and minimum value, and confirmed this empirically. These explorations hence opened a window onto the functioning of the `random` tool – each of the numbers within the defined range had an equiprobable chance of being selected, which meant the mean value of the $n - 1$ randomly chosen values tended to the midpoint of the range. The larger the data set, the greater this tendency and the more 'equal' (uniform) the spread of numbers selected. This made sense to them, but for Edith raised a new problem. She was worried that, in practice, families with 8 members are not as just as likely as families with 4. This raised a real dilemma for her: was the `random` tool, which draws numbers from a uniform distribution, a sensible tool to use in this situation?

THE LAW OF SMALL (!) NUMBERS

A second group developed a strategy which involved generating sets of size n in which each value is selected at random from the specified range and then the sum of the n values calculated. While the sum of the values is not equal to n times the given mean, the data set is discarded and a new set generated. This process continues until a data set with the required mean is obtained. This is clearly a computer-mediated solution strategy; it seems unlikely that it would emerge in the paper-and-pencil setting, as it involves a considerable amount of redundant calculations. As it turned out, it had some limitations in the computer context as well, limitations that led its proponents to discuss notions relates to probability.

It is a strategy that was relatively easy to program using the Imagine tools, the problem was that, on occasions, it was extremely slow to return a result – so slow that the group members tended to give up on it and interrupt the process before a result was obtained. The problem was not the code itself: the procedure worked fine, regardless of the value of the mean, as long as both n and the interval between the minimum and maximum value were pretty small, but for larger values of n, unless the given mean was a value close to the midpoint between minimum and maximum, the procedure took ages to find an appropriate data set. To explain this delay, like the members of the first group, those responsible for this strategy were stimulated to reflect upon the random function. They modified the procedure so that all the data sets and their totals were shown while it was running. The group's reporter summarised the results of this investigation:

> *There are more possible data sets whose sums are close to the 'middle' value times the number* n. *When the mean is close to the 'middle,' there are more data sets that fit so a correct set is found without too much delay. If the mean is very close to the maximum or minimum, there are not many possible sets that work, you need lots of the same number and the computer does not seem to find them easily. For example, if you have a data set with a minimum of two and the mean is also two, then only one data set is possible, which consists entirely of twos. In sets made up of only a few values not so many different sets exist so the procedure works OK.*

To overcome the problem described, Ivanildo asked Melanie (the programmer of the group) if it would be possible that instead of building the data set up in one go, the size of the set could be broken down into smaller sets. If this was possible, he reasoned, then data sets with the correct mean could be generated for these smaller sets and finally all the data sets joined together. In coming up with this solution, he expressed a particular property of distributions that not all of the sub-group members were initially sure about: a set formed by joining two or more sets with the

same mean, will also have this mean. Following Ivanildo's suggestion, Melanie's original procedure was modified so that the given *n* was broken down into sets of five plus any remainder (using the Imagine tools `div` and `mod`). The group were well-satisfied with the result and convinced that Ivanildo's reasoning was valid. Figure 2 shows two distributions generated by the final version of their procedure.

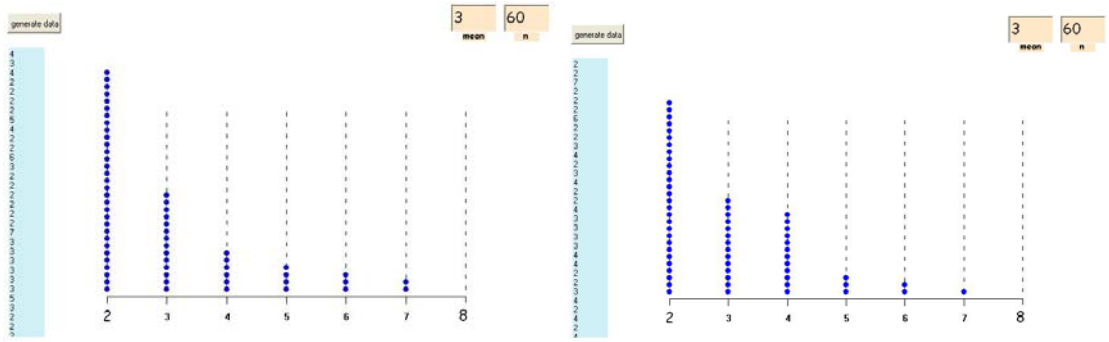

Figure 2: Two data sets of 60 values between 2 and 8 and a mean of 3

FROM MEANS AND SPREAD TO DISTRIBUTION AS A SPACE OF POSSIBLE VALUES

The challenge of simulating data sets given the mean and the range of possible values certainly involved the participants in deep thinking about random processes and in making various generalisations about how the shape of a distribution is related to these properties. This points to the value of what has been termed a *bottom-up* approach to designing tools appropriate for the learning of statistics (Bielher, 1997; Konold, 2002). *Top-down* tools are based upon expert practice and derive from software for professional statisticians, essentially providing access to a subset of the conventional plots and measures available in their more sophisticated parents. From Konold's point of view, the *top-down* approach can have the effect of emphasising the learning of particular means for representing and summarising data over and above the expression and exploration of the concepts that underlie them – such as spread, variation, centre and shape. In contrast, *bottom-up* tools have their basis in the learners' practices and reasoning. By involving participants in this project in the design of their own tools, these concepts came to the centre stage and various features of distribution were emphasised. In addition, it proved necessary to explore the functioning of the particular tool for generating random numbers that was available in the software used.

A major concern that emerged in the first of the two groups described in this paper was the fact that the random tool that was being used in their simulations selected values with an equal probability, while the distributions that they were seeking to construct were not uniform. This problem seemed especially evident to them because of the way the challenge was originally expressed in terms of distributions of family sizes. While at the moment of programming this context may have been left completely aside, at the point of assessing the procedures that were constructed, it came back to the forefront creating a conflict between distributions that did not correspond to their expectations, but were mathematically valid in terms of the constraints of the challenge. The second group focused much less on the problem of family sizes, happily working with sets in which all families were of the same minimum size – and even setting this size to 0 as they let go completely of the context.

One interesting aspect of the work of this second group, as presented in the reporter' description presented above, is that in order to understand why their procedure was not always efficient, the group was provoked to begin to think about the space of possible values of a distribution with particular properties – a concept central to thinking about distributions theoretically. Listening to the second group's report seemed to help Edith, one of the first group members, clarify further her problem with the random tool:

> *I don't know if this is right, but with random, the more we picked the more equal the spread of numbers and I don't think we want an equal spread. It should depend, depend on … on the*

situation, yes, but on the mean as well. If we find all the right lists and then choose from them, not like at random, well maybe a different random, the set might be more realistic.

Edith now conjectures that the 'ideal' shape of a distribution can be found by constructing a list which contains all the possible data sets for a given mean and range, then plotting all the values in this list. She has proposed this as a new challenge to the group.

REFERENCES

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review,* 65, 167-189.

Costa, N (2004). *Formação de Professores para o Ensino da Matemática com Informática Integrada a Pratica Pedagógica: Exploração e Análise de Dados em Bancos Computacionais*. Doctoral thesis - Pontifícia Universidade Católica de São Paulo.

Hancock, C., Kaput, J. and Goldsmith, L.T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist*, 27, 337-364.

Konold, C. (2002). Teaching concepts rather than conventions. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics*, Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Mokros, J. and Russell, S.J. (1995). Children's concepts of average and representativeness. *Journal for research in Mathematics Education*, 26(1), 20-39

Papert, S. (1991). Situating constructionism. In I. Harel and S. Papert (Eds.), *Constructionism*, (pp. 1-11). Norwood, NJ: Ablex Publishing Corporation.

Santos, S. (2003) *A Formação do Professor Não Especialista em Conceitos Elementares do Bloco Tratamento da Informação: Um Estudo de Caso No Ambiente Computacional*. Masters dissertation - Pontifícia Universidade Católica de São Paulo.

Stella, C. A. (2003) *Um Estudo Sobre o Conceito de Média, No Contexto Brasileiro, Com Alunos da 3a Série do Ensino Médio*. Masters dissertation - Pontifícia Universidade Católica de São Paulo.