From: Blum, W./Niss, M./Huntley, I.: Modelling, Applications and Applied Problem Solving: Teaching Mathematics in a Real Context. Chichester: Ellis Horwood 1989, 123 - 130

CHAPTER 15

Enhancing Probability Education with Computer Supported Data Analysis

R. Biehler
Institut für Didaktik der Mathematik (IDM), Universität Bielefeld, FR Germany

ABSRACT
Computers can be used in applied mathematics and science teaching to support modelling, simulation and the analysis of empirical data. These possibilities are analysed critically with regard to probability education at school level where data-free modelling is predominant. An initial impact of computers may be an increasing gap between probability and statistical data analysis, if case simulation is used in a limited sense. Some perspectives for using data analysis and theoretical modelling as complementary approaches will be developed. The use of models as reference frames plays an important role.

1. INTRODUCTION: MODELLING AND REAL DATA IN PROBABILITY AND STATISTICS

Combining the collection and analysis of real data with building theoretical models is a major opportunity of using computers in applied mathematics and science teaching (see eg Barclay 1988). It can be used to support students in actually *doing* science instead of a continued use of classical word problems, which also undermine an adequate understanding of the model concept. For probability education at lower and upper secondary school level, the relation between modelling and data analysis under the new technological conditions has to be explored further. Some thoughts on this will be developed in this paper.

A comprehensive approach to modelling in probability was seldom put into practice. This is partly due to complexity: a model for describing free fall is extremely simple as compared to a description for the complex structure of data in the most simple random experiment, for example coin tossing. In contrast to the free fall, people usually do not have experience with a long series of coin tossing. Moreover, comparison between model and data is often done informally in science (education). Comparison of probability models with real data is more complicated and the object of a full scale theory (statistics). This situation is reflected in the division of labour at the scientific level. Even if we look at a very model-conscious university course like that developed by Breiman (1969, 1973), we find an instructive separation. In the probability part, data-free modelling is practiced. Models are constructed relying on basic heuristics concerning the physical meaning of equally likely outcomes and of independent trials. This basic raw material is used to construct complex and compound probability models. In the statistics part, *initial* models are assumed to be given, and statistical methods are designed for this situation. A more or less *data-free* statistics is practised. A certain exception is testing goodness-of-fit, where the initial model is put into question. Breiman (1973, page 216) gives as a basic heuristic principle : "It is the departures from the hypothesized distributions that frequently give them most informative and interesting insights into the nature of the physical system producing the data". This attitude of using probability models as *reference frames* is also made explicit by Feller (1968).

On a limited scale, there is quite a tradition at school level in attempting to integrate ideas from statistics and model building into an (applied) probability course from the beginning, instead of reproducing the division of labour at the university level. All the activities offered to children with simple random devices such as games, spinners, dice and coin can fall into this comprehensive or holistic approach and comprise activities in all the areas. Working with random devices has pedagogical and practical significance because it provides opportunities for practical and cooperative work. It has epistemological significance because it supports developing the probability concept as a tool for handling random situations in real life.

2. THE INITIAL IMPACT OF COMPUTERS

Although the computer might have been used for prolonging the latter approach, it had certain effects in the opposite direction. For probability education, the use of simulation is an obvious extension, which is already supported by a very limited technological equipment such as programmable hand-held calculators. Many practical suggestions for use in the classroom partly widen the gap between probability and statistical data analysis because:

(a) simulation is used to foster a sequential and dynamic interpretation of change experiments as *stochastic processes*, in contrast to the point of view of statistics, which prefers finite samples for inference and decision;

(b) simulation can be attractively used to extend the data-free modelling beyond the simple situations of current school mathematics, for instance including Markov processes or more complicated problems concerning simple situations (for example waiting times), which is mainly due to the assumption robustness of computational models;

(c) simulation is suggested for use as a partial replacement of real data, for instance by simulated Galton Boards and visualisations of the law of large number with artificial data (see also Biehler 1988b, 1989).

On the statistics side, there are complementary indications of further separation. Statistics liberated itself, with the help of computers, from the universal control of probability and developed model and probability-free approaches, mainly for Exploratory Data Analysis (EDA). This has led to several novel approaches at the school level that aim at providing more independent room for data analysis instead of functionalising descriptive statistics for purposes of learning and teaching probability (see eg Biehler 1988a). In practice, however, the methods of data analysis are also applied to improve model-based statistical analysis, ie to overcome the status that too unrealistic and grossly misleading models are used in statistics. Data analysis can be used to develop initial models for a situation, to check model assumptions, and to analyse the structure of deviations (residuals) from a model used as *reference frame*.

Although these relations exist, data analysis and modelling can be regarded as different and partly complementary strategies for understanding a situation: a model may have a certain predictive power and may provide a simplified explanation and basic insight without being able to explain everything in the data. Data analysis may reveal many peculiarities of interest without being able to provide such general insights as a model might do. There are situations where only data analysis is possible, in other situations no data are available and modelling is a last resort. This distinction between the level of describing relationships and the level of explaining them by theoretical models is common in science, but not in probability.

Simulation can play important roles.

(a) The extended modelling repertoire can be exploited to develop theoretical models for complex real data and study them by simulation. The comparison between model and real data could then be done by comparing simulated and real data;

(b) Data and model can be compared with simple graphical methods as developed by EDA. If a more theoretical approach seems to be necessary in order to be able to judge whether a deviation could have been produced by chance, there is new room for developing ad hoc criteria and studying their characteristics under the relevant probability assumptions by simulation. This approach is related to BOOTSTRAP methods.

This new flexibility could support a similar flexible and complementary approach toward modelling and data analysis at the school level. A more or less normative modelling cycle, which gives each element its place, may become very misleading.

3. PERSPECTIVES AND EXAMPLES FOR USING REAL DATA IN PROBABILITY EDUCATION

The role of real data in individual applications of modelling and their role in long term development of probability in the school curriculum are two distinct problems. For instance, it could be decided to offer a course in Exploring Data and to rely on the manifold experiences made available in such a course when probability is introduced later without practising real data anlaysis so much at that stage. In the following, some basic options will be discussed.

3.1 Exploring and describing randomness

The new technological possibilities for exploring data could help to introduce the probability concept as a *theoretical* concept (for this notion, see Steinbring 1989) which is the basis of a theory which aims at describing and understanding empirical phenomena. We may illustrate this with R. von Mises, who intended to rebuild probability theory as a kind of empirical theory similar to mechanics. The objects of probability theory were conceived as 'collectives' that can be characterised through two fundamental empirical phenomena (Urphänomene) : stablising frequencies and the principle of the excluded gambling system. The latter is closely related to the concept of stochastic independence and expresses the complex structure in sequences of real random numbers. There is a direct line from von Mises' foundational efforts to the modern theory of random numbers. On the other hand, school probability has always *suppressed the complexity* in sequences of data favouring the stablisation of frequencies. Also, students will usually not have experienced the many different manifestations of laws of large numbers in reality, rather they will at best have seen one graph with a make-believe sequence of real data. Now (pseudo-)random numbers and their structure have entered the neighbourhood of the school curriculum together with simulation. However, if the laws of large number shall be established as a thought in reality, the first thing would be to use computers to explore data bases containing real data from different

subject matter domains (insurances, vital statistics, etc) in order to support a rich knowledge of the generalised laws of large number. More generally, or philosophically, laws of large number can be interpreted as the existence of statistical regularities for a large number of cases despite irregularities and unpredictability in individual cases, as 'order emerging from chaos'. Experience with exploring data, where summarising, smoothing or middling data help reveal structure, can be related to this topic.

3.2 Modelling and reproducing variation in data

In the context of games of chance, the perspective that probability models are constructed to reproduce variation in data is not a natural one. Historically, it was the problem of measurement error that was a new challenge to develop probability models that were able to reproduce observed variation. This was more or less the beginning of a probabilistic research programme which aimed at conquering all kinds of processes with variation and uncertainty by using probabilistic methods.

In schools, using the Galton Board as a starting point for the development of probability lies on these lines. With computer support an extension would be possible. For instance, having more experience with real data on measurements, comparing the quality of different instruments for measurement (different variation), experiencing the existence of outliers and so on. Complementary, simple error generating probability models which support the idea that the observed error is the sum of many small independent errors could be actually studied by simulation, and the simulated data could be compared with data from real measurements. This would provide quite a different context than merely illustrating the mathematical central limit theorem by simulation. If students have had experience with data analysis concerning structures in bivariate data, in time series data, in sequential data (eg letters in texts) and the like, these may be taken up and be partly reproduced by simulation.

3.3 Modelling and data analysis as complementary strategies

If a library of data sets is available, word problems in probability could be enhanced with a more realistic context. A typical word problem in probability is the birthday problem: if s people meet in a room, what is the probability p(s) that at least two of them have their birthday on the same day of the year? p(s) can be computed on the basis of the assumption that each birthday is equally likely. The well known but surprising result that $p(s) > 0.50$ is true for surprisingly small s can be explained by combinatorical thinking. In a computer context, the birthday problem could be used as an opportunity for exploring how birthdays are really distributed over the days of a year. This could be extended to a small project in EDA. Smoothing or summarising data by month may help the analysis as well as plotting the residuals, ie the

deviation between the relative frequencies and the expected 1/365. Results of several simulations of a uniform distribution could be graphically compared to the pattern in the real data, and p(s) could be calculated on the basis of a simulation taking the real distribution as the theoretical one. In the end one may have learned something new about patterns of human births as well as about the value of simplified models for explaining qualitative results, ie the high probability for a coincidence.

3.4 Comparison of model and real data

There can be many arguments for making a theoretical model experiential for students through simulation (see Biehler 1988a, 1988b). For instance, an interesting environment is provided by the program COINTOSSER INVENTION ANALYSER for the collection PROBABILITY AND STATISTICS PROGRAMMES. The basic idea is to use nine coin tossing machines that are implemented on a computer but where the models are hidden from the user. Each machine is either a fair coin tosser or represents a certain deviation from the ideal model of equal chance and independent trials (different kinds of dependencies, changing probability for HEAD, deterministic patterns, unequal chances for HEAD and TAIL, but independence). Students have several options for analysis and display of data for exploring the different machines and learning about what structure can be expected in random sequences. Such knowledge on theoretical models can then be used as a theoretical perspective for exploring real data. There are many interesting applications where the use of the model as a reference frame gives new insight into subject matter problems, for instance in process and quality control the expected structure of random series is an important reference frame.

If data libraries were available in probability education, this would be a rich source for showing students what kind of situations can be modelled by the standard distributions of the curriculum like binomial, Poisson and normal distribution. Nevertheless, the use of a reference frame may be the source of many more interesting applications, for instance the deviation from the binomial model concerning the distribution of male children in families with 5, 6, 7, ... children, how the balls in Galton Boards are really distributed and so on. As has been pointed out at the beginning, graphical methods may be of considerable help here, as well as simulations, which can be used for getting an idea on how large tolerable chance variations may be.

4. SOFTWARE REQUIREMENTS

Although there exist several software systems that support the modelling of dynamic systems and that would be usable in secondary education, such systems are lacking for the specific purposes of probability modelling. At present, array-oriented interactive languages like APL or

PC-ISP would probably have enough flexibility and options for graphical display, but they present problems with regard to user friendliness and ease of use. It has to be further explored whether their extensibility and adaptability can be used to construct simpler learning and application environments for the purposes described above. Statistical packages have to be carefully examined to see whether they have spreadsheet functions and flexible options for generating random numbers that may also support modelling, simulation and data analysis. Such research is currently done in the SOMA (softwaretools for mathematics education) project at our institute.

REFERENCES

Barclay, T. (1988). MBL to model: Combining real world data with theoretical models. In Blum et al 1988.

Biehler, R. (1988a). Educational Perspectives on Exploratory Data Analysis. In Morris, R. (ed) *Studies in Mathematics Education Vol 7: Teaching of Statistics*. Paris, UNESCO.

Biehler, R. (1988b). Computer Simulation as Tool and Object of Teaching and Learning Probability and Statistics. In Blum et al 1988.

Biehler, R. (1989). Computers in Probability Education. To be published as a chapter in Kapadia 1989.

Blum, W., Berry, J., Biehler, R., Huntley, I., Kaiser-Messmer, G. and Profke, L. (eds) (1988). *Applications and Modelling in Learning and Teaching Mathematics*. Chichester: Ellis Horwood.

Breiman, L. (1969). Probability and Stochastic Processes: With a View Towards Applications. Boston: Houghton Mifflin.

Breiman, L. (1973). Statistics: With a View Toward Applications. Boston: Houghton Mifflin.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. New York: J Wiley and Sons.

Kapadia, R. (ed) (1989). Chance Encounters. *Probability in Education: A Review of Research and Pedagogical Perspectives on Probability in Education*. Dordrecht: Reidel, in preparation.

PC-ISP: Interactive Scientific Processor. Artemis Systems. Inc New York and London: Chapman and Hall; for IBM-PC and comp.

Probability and Statistics Program (1986). Green, D. et al: Capital Media c/o ILECC, John Ruskin Street, London, SE5 0P2, BBC-micro.

Steinbring, H. (1989). The theoretical nature of probability and how to cope with it in the classroom. To be published as a chapter in Kapadia 1989.