MODELING AND SIMULATIONS IN STATISTICS EDUCATION

Brigitte Chaput[1], Jean Claude Girard[2] and Michel Henry[3]
[1] École Nationale de Formation Agronomique, France
[2] Université Claude Bernard Lyon 1, France
[3] Université de Franche-Comté, France
brigitte.chaput@educagri.fr

In France, recent mathematics curricula reinforce the teaching of statistics and probability. They recommend starting with an experimental approach introducing the observation of sampling fluctuations and the construction of random experiment simulations in order to prepare students for theory. This approach raises the problem of the didactical practice of random experiment modeling and simulations.

INTRODUCTION

In France, for almost 40 years, the IREM network (26 Research Institutes in Mathematics Education) has followed the evolution of the curricula. The aim of each IREM is to bring help to teachers by giving them background knowledge and tools to teach. The work of each IREM is gathered within national commissions. One of these Inter-IREM commissions deals with statistics and probability teaching. It was created in 1991 on the occasion of a curricula change applied to the *lycée* baccalaureate classes. Indeed, the 1991 national curricula required teachers to present the notion of probability in a frequentist approach while, up to then, the curricula only prescribed a combinatory introduction to probability. At the time, the aim of our commission was to contribute to the teachers' reflection and to examine the impact of this approach on the students' understanding of the basic notions. The curricula newly introduced in 2000-2002 require implementing an experimental approach to statistics and probability through simulations. The purpose of this introduction is to prepare students to understand the law of large numbers and to grasp the interaction between the notions of relative frequency and probability. In the mathematics curricula, the statistics section is concerned with simulation and sampling fluctuation as well as with the basic notions. The orientations concerning simulation and sampling fluctuation are presented in Table 1.

Table 1. Statistics section of the French mathematics curricula (*lycée* first year)

CONTENTS	EXPECTED SKILLS	COMMENTS
• Definition of the distribution of relative frequencies in the case of statistical series taking a small number of values. Definition of the relative frequency of an event. • Simulation and sampling fluctuation	Designing and applying simple simulations from random digit samples.	The calculator "random" key may be presented as a process giving a list of *n* digits (the decimal part of the displayed number). When the process is replicated a large number of times, the series obtained will be out of order and without periodicity, and the relative frequencies of the ten digits will be more or less equal. Each student will produce size n simulations (*n* ranging from 10 to 100 according to the case) with the calculator; the simulations can be gathered into one or several size *N* simulations after considering the variability of each result. The teacher will then be allowed to provide already prepared size *N* simulations obtained with computers.

As the teaching of statistical tools was reinforced and simulations were introduced, the Inter-IREM commission felt the need to clarify the status of probability as a part of teaching statistics and the role of probability in learning modeling, where probability is considered as a model for data (statistics). This article addresses a synthesis of the work of the commission dealing with an analysis of problems involved in introducing modeling and simulation in the teaching of statistics in line with the new instructions and reflects on the consequences on the training of teachers.

 In C. Batanero, G. Burrill, C. Reading & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference.*

NOTION OF *MODEL*

In the sixties, the necessity of a tool within statistics that would no longer be a mere description of reality but serve to make predictions gave birth to the notion of model that contains theoretical knowledge allowing the user to evaluate, to interpret and to generalize reality. The commission accepted this view of a model:

> A model is an abstract, simplified and idealized representation of a real object, a system of relations or an evolutionary process, within a description of reality. (Henry, 2001, p. 151)

The presentation of a probability model to students in the statistics classroom accounts for various levels of abstraction and formalism. First for a didactical purpose, some basic models may be presented in relation to reality using everyday terminology. Then, the objects from reality are idealised by selection of relevant characteristic properties. In the process, one obtains the so-called pseudo-concrete models. An example in probability is the Urn Model: in a first step a real urn containing balls (indiscernible to the touch...) is represented by an ideal urn which is a probabilistic model where the implicit hypothesis is the equiprobability of the balls in a random draw. Then, among the different types of representations, the mathematical language and the mathematical symbolism (e.g. the Bernoulli model) allow strong descriptions that contain general properties and algorithms with which we can operate. We will call these representations mathematical models. A problem is that they are often so familiar that one is inclined to confuse mathematical models with the related ideal objects, which, in their turn, are often confused with the reality they model.

STEPS OF MODELING PROPOSED IN THE TEACHING OF STATISTICS

In the didactic analysis of the modeling process, in particular in probability-statistics we distinguish three steps: pseudo concrete model, mathematization and validation.

Reality description and pseudo-concrete model

The first step consists of the observation of the concrete situation and the description of it in usual terms. The description is controlled by a so-called theoretical approach, i.e., a scientific knowledge based on pre-designed general models in order to evaluate what is relevant. Students are required to translate the description into a simplified and structured system, to choose the characteristic aspects of the real objects in order to design the relevant pseudo-concrete model. From a didactical point of view, this stage is called the contextualization of previous knowledge. Then, the work hypotheses are set out to describe the situation; for instance in the Urn Model, balls are assumed to have the same chance of being drawn. The experimental process also consists in acting on the reality in order to study the evolutions and the invariants. It requires building up an experimental protocol, i.e., a set of instructions to be followed in order to carry out the experiment and to reproduce it, if necessary.

Mathematization-formalization

Then, comes the second step: mathematization. The work hypotheses lead to the model hypotheses; for instance, in the Urn Model, a uniform distribution can be used to model the situation. Students have to be able to represent the model in a suitable mathematical symbolic system, to translate the question asked into a purely mathematical problem. Then the model must be formalized and its hypotheses checked. Finally, students have to choose the right tools appropriate to solve the abstract problem.

Validation

The third step consists first in translating the mathematical results according to the previous pseudo-concrete model, then in giving meaning to the mathematical results in order to create answers to the original question, and then again in confronting these answers to the

model hypotheses. Lastly the answers have to be put into perspective in order to estimate their validity. These different stages may require specific training in other domains. In some instances, this supposes a specialized knowledge of the studied situation and no longer concerns a mathematician. A domain specialist will be able to validate the conclusions according to his or her knowledge of the situation.

DIFFICULTIES LINKED TO THE PROCESS OF MODELING

The French curricula suggest for statistics an approach similar to the process of constructing the Euclidean model in geometry, whose starting point is the observation of real objects. In geometry, the objects are discovered globally, their properties are progressively drawn, and finally the mathematical objects are defined. This conceptual jump generates difficulties for the students who gradually find out about the scientific process. Compared to geometry, statistics and probability theory is taught in a different context. The teaching of geometry starts in primary education and lasts ten years (which leaves no room for a conceptual jump), while the teaching of statistics and probability occurs only in the first two years of the *lycée* when students have reached the age of 15 to 16 years, and naïve conceptions have already settled in their minds. Particularly the perception of randomness is not univocal and is linked to many different beliefs.

Another difference is that most geometry problems are posed in the Euclidean model, which is rarely used to solve concrete problems; at best it concerns pseudo-concrete problems, and often the associated modeling is already completely detailed for students. Contrarily in almost all the probability problems, the stage of modeling is present with a concrete approach, and the contexts are close enough to the learning situations involved so the students can transfer from the concrete situations to the usual probability laws. As the learning period is very short, the students may not make sense of this approach and may not have a good learning of probability modeling.

SIMULATION AND MODELING

The curriculum of statistics in the first year of the *lycée* (students aged 15-16) introduces simulations; however, teachers interpret simulation as only requiring students to represent the outcomes of a concrete experiment. This restricted acceptation neither raises the problem of the subjacent theoretical model necessary to solve the task nor reveals the absence of such a model in the students' minds. For example, in simple situations such as throwing die, equiprobability is implicitly accepted and associated with the uniform discrete distribution that is supposed to control the random digits; however, this distribution has not been first taught to the students; yet designing a simulation requires a minimal knowledge of probability that students do not actually have. The point is how to justify to students the equivalence of real or pseudo-concrete random experiments with a computer simulation, judiciously programmed from a theoretical model. The equivalence is ensured by the fact that both experiments are relative to the same probabilistic model, a concept not yet available to the students. Another point is how to interpret the sampling fluctuations observed in the repetition of the simulated experiment. Without answers to these questions, teachers are in a difficult didactical situation.

This didactic inconsistency is pointed out in the document accompanying the curricula, which gives teaching tips for the *lycée* second year program (students aged 16-17):

> The respective positions of modeling and simulation will be briefly clarified: modeling consists of associating a model with experimental data while simulating consists of producing data from a defined model. The simulation of a distribution P will be presented; a simulation with a random digit table can only be done if P can be built as image of an equidistributed distribution. In order to simulate an experiment, it is necessary to first associate a model with the current experiment and then the model distribution is simulated. These stages may be detailed... (GEPS, 2002, p. 72)

This very pertinent comment specifies through an example the meaning of model, but it

cannot avoid a reference to the theoretical concept of distribution. Thus, in the document, the didactical importance of a theoretical support for conceptualization and ultimately for the acquisition of a real scientific knowledge is emphasized. In the same way, the document provides enlightenment on modeling:

> Modeling a random experiment consists of associating the experiment with a distribution of the set of the possible issues. The experiment modeled leads to choices generally delicate to make, except in certain cases when considerations appropriate for the experimental protocol suggest an a priori model. It occurs for instance in the dice or coin throwing where symmetry considerations lead to an equidistributed probability law. But the experiment of reference should be treated avoiding general talking about what is modeling or what is not. (GEPS, 2002, p. 70)

In order not to consider modeling as an aim in itself, safeguards are given:

> Apart from such cases when considerations linked to the experimental nature suggest a model, the model choice from experimental data is much more delicate and will not be tackled in the secondary school education cycle. If necessary a model may be given indicating that statistical techniques have allowed determining and validating such a model. (GEPS, 2002, p. 70)

Thus the link with the statistical data is not skipped, according to the experimentalist option of the curricula. It reappears once again, later in the document:

> Modeling doesn't belong to the right or wrong logic: a model is neither right nor wrong; it may be validated or rejected on the basis of experimental data. One of the first functions of the so-called inferential statistics is to associate a model or a range of models, to a random experiment and to specify procedures for validating the model. These models have to be appropriate to the experimental data and adequacy has to be justified... In order to determine and/or validate a probabilistic model, the first available tool is a mathematical theorem called the law of large numbers. An intuitive wording of the theorem is: in the theoretical context defined by a distribution P on a space E, in a series of n identical and independent experiments, the relative frequencies of the elements of E tend to their probabilities as n increases indefinitely. (GEPS, 2002, p. 71)

In the approach suggested in this document, giving meaning to the basic probabilistic concepts such as probability, distribution, random variable, expected value and standard deviation, thanks to statistical observation, the law of large numbers will play a decisive part. The computer helps to prepare understanding of this law as it allows working quickly on a large amount of statistical data. Yet, using a computer for its mere power and speed for the sake of presenting a large rich set of new random experiments is not satisfactory. The didactic interest of simulation lies elsewhere, in the analysis of the random situations, the design of model hypotheses and the translation of them into computer instructions that are necessary previous to simulations; once this is done, computers can be used to find solutions to problems that may not even be solved by calculations. Although computers can only show the equidistribution of random digits they generate, their use in school as pseudo-random digit generators eases the understanding by the students of the notions of relative frequency, sampling fluctuations and probability.

FREQUENTIST APPROACH TO PROBABILITY AND THE LAW OF LARGE NUMBERS

The stabilization of relative frequencies when the number of experiments increases is an observed fact and is classified among the random laws in the physics meaning of the term law. From the definition of relative frequency, it is obvious that when tossing a coin for the $1{,}000^{th}$ time, the result obtained will have far less effect on the relative frequency of heads than when tossing the coin for the 10^{th} time. This phenomenon has been known since the ancient times and

for centuries allowed players to estimate their stakes and organize their bets without, however, resulting automatically in the construction of the concept of probability, which only appeared in the second half of the 17th century, after the notion of winning expectation.

However, the definition of probability as a stabilized relative frequency (after the French curricula prescription concerning this frequency limit) raises serious epistemological problems because it characterizes a mathematical – and consequently abstract – object (probability) from experimental data (frequency). Reality and mathematics domain could be confused. However, the great probability educationist Alfred Rényi, accepted this point of view:

> The probability of an event will be the number around which the relative frequency of the considered event fluctuates... Thus the probability is considered as a value independent of the observer; it indicates approximately the relative frequency of the considered event in a long series of experiments. (Rényi, 1966, p. 25)

After the demonstration of the law of large numbers, Rényi pointed out the adequacy of the probability theory:

> The relative frequency stability has been demonstrated mathematically. It is amazing that theory makes a precise description of this stability possible; that shows without any doubt its power. (Rényi, 1966, p. 144)

Then he tried to explain what appears to be a vicious circle:

> Indeed we have defined probability using relative frequency stability. In reality, there are two different issues. The definition of probability as a value around which the relative frequency fluctuates is not a mathematical definition but a description of the factual fabric of the notion of probability. On the contrary, the Bernoulli law of large numbers is based on a mathematical definition of probability, and therefore, there is no vicious circle. (Rényi, 1966, p. 144)

The mathematical definition mentioned by Rényi is the definition of a measure on an abstract set. Thus the demonstration remains within the mathematical model, and the law of large numbers is a mathematical theorem. The assimilation of the theorem to the phenomenon of the relative frequency stabilization would come close to an epistemological confusion. This epistemological difficulty, due to the frequentist approach, has often been mentioned, and teachers should be careful not to keep this type of contradiction when they start introducing the probability concept in connection with modeling. A meticulous wording of the law of large numbers, even in the simplest form of the Bernoulli theorem, presupposes a mathematical definition of probability and should not be introduced in a confusion context between model and reality. This definition can be based on the equiprobability for a finite set of cases. This hypothesis may seem to be restrictive. It can also generate didactical obstacles, and it is better to describe it only in the situations where this model hypothesis is obvious. The formulation itself generates understanding difficulties. Here is Bernoulli's simplest form of the theorem:

> During the indefinite repetition of a Bernoulli experiment (i.e. with two issues: success with a probability p and failure with 1 – p), the probability that the difference between p and the relative frequency of the successes obtained in n experiments is higher than a given ε, tend to 0 as n increases indefinitely. (Girard, Henry, 2005, p. 156)

In the above sentence, there are two probabilities that are quite different from the conceptual point of view (Laplace used two words probability and possibility):

- The (objective) probability p of a success in the Bernoulli experiment, for instance the proportion of white balls in a Bernoulli urn,
- The probability $P(|f - p| > \varepsilon)$ that can be considered as a (subjective) control of the

experimental data, which establishes links between the observed relative frequency *f* and the theoretical probability *p* of success.

This wording raises an epistemological obstacle because it seems to include both a real and a theoretical object in the same formula. But here, the relative frequency *f* comes from a binomial variable defined on the probabilized space in order to represent the proportion of success obtained among *n* experiments. So, it appears as a model object, and the theorem is a consequence of the properties of the binomial coefficients that occur in the law of the variable.

The lack of a shared definition of probability with the epistemological difficulty mentioned above explains why the *lycée* second year curricula do not introduce an explicit wording for the law of large numbers. As suggested in the accompanying document, using the theorem to validate a model associated with a real or simulated random experiment, appears as the key of the modeling approach, and the curricula authors have chosen a simplified formulation. Indeed, without the expression of $P(|f - p| > \varepsilon)$, the theorem remains little operative and has a qualitative part that appeals to intuition.

If it is right that "statistics thinking appears with the awareness of sample fluctuation", does the observation of the relative frequencies distribution fluctuation during several simulations of the same random experiment, lead to the acceptance of the idea of a theoretical law linked to the experiment? The learning of probabilities with the law of the large numbers and simulations is not that simple, and an earlier start and a longer learning period are required.

CONCLUSION

Teachers should be aware that, unlike other mathematical notions, statistics and probability knowledge is rooted in everyday life. Understanding the probabilistic modeling process is an essential stage, but the introduction of probabilistic notions poses specific problems, so that students come across a new difficulty when they have to link the probabilistic notions to reality. The probability theory taught in a finite context, as required by our curricula, is very simple, but its abstract model part is not direct, particularly in the somewhat artificial situations presented in school. Modeling is a critical stage in the use of the probability theory, especially in the different statistics fields.

The construction of mental images relative to randomness is delicate. It is necessary to present activities to the students before the *lycée* in order to create such images. Recent research has shown that younger students could use simulation to build equivalent experiments and that they could model situations by analogy with Bernoulli urns. The next French *collège* (secondary school up to age 16) curricula to be implemented in September 2008, introduces an initiation to the probability concept, so that it will be possible to establish links between statistics and probability for two additional years. As for geometry, the sequence could be: observation, building, reproduction, description and representation of random experiments. The techniques of descriptive statistics at *collège* (means, percentages, bar charts, box plots, stem and leaf plots...) will be used to communicate and sum up the results of these experiments. The notions of sampling fluctuations and model could then be progressively constructed.

REFERENCES

GEPS (2002). Document d'accompagnement des programmes de lycée, mathématiques (Documents accompanying the secondary programmes, mathematics). In C. Robert (Coord.), *Direction de l'enseignement scolaire*. [CD-ROM] MEN, Paris, France: CNDP.

Girard, J. C., & Henry, M. (2005). Modélisation et simulation en classe, quel statut didactique? (Modelling and simulation in the classroom, what didactic status?). In Commission Inter-IREM Statistique et Probabilités, *Statistique au lycée,* APMEP vol. 156 (pp. 147-159). France: APMEP.

Henry, M. (2001). Notion de modèle et modélisation dans l'enseignement. In Commission Inter-IREM Statistique et Probabilités, *Autour de la modélisation en probabilités* (pp. 149-159), Besançon, France: PUFC.

Rényi, A. (1966). *Calcul des probabilités,* Dunod. France: Paris. Reed. (1992) Jacques Gabay.