



HAL
open science

Using spreadsheets to teach probability in French high school

Bernard Parzysz

► **To cite this version:**

Bernard Parzysz. Using spreadsheets to teach probability in French high school. CERME 10, Feb 2017, Dublin, Ireland. hal-01927774

HAL Id: hal-01927774

<https://hal.science/hal-01927774v1>

Submitted on 20 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Using spreadsheets to teach probability in French high school

Bernard Parzysz

Université d'Orléans & L.D.A.R. (université Paris-Diderot)

parzysz.bernard@wanadoo.fr

Including a 'frequentist' point of view has resulted in experimentation becoming an important issue in the teaching of probability in high school. Spreadsheets are now widely used, but the status of the results produced and how to use them are not always clear for the students, since two domains are at play in turn: statistics and probability. Through the French example – but this can also be applied to the teaching of probability in other countries – this paper reviews some questions about spreadsheets, namely simulating random experiments and shifting from discrete to continuous distributions.

Keywords. Probability, simulation, continuous distribution, spreadsheet.

In France as in many other countries, probability has become a prominent subject in the teaching of mathematics, in link with the growing importance of numerical data in everyday and professional life and the development of technologies allowing to process them. Conversely, the point of view on probability has evolved, taking into account a 'frequentist' point of view implying students making experiments. Among the technological tools spreadsheet has become much used in the teaching of probability, the main reason being that it makes getting a large number of tries of a random experiment very easy and fast. This paper is based on the French current situation, but it certainly can apply to many other countries all over the world. Its aim is to give some insights on how new didactic questions have occurred and have now to be tackled by teachers in their classes. As it has already been noticed, "it is not good enough to only consider which technology to use, but (...), in order for effective learning to take place, it is how the technology is integrated into the curriculum and learning process and how the teacher uses it that are vital" (Pratt, Davies & Connor, 2008, p. 98), the more so that "most teachers have little experience with probability and share with their students a variety of probabilistic misconceptions" (Batanero et al., 2005, p.1). I shall discuss some questions about the use of spreadsheet in high school, namely:

- the nature of simulation, implying using a model of the random experiment and making in turn intervene probabilistic and statistical paradigms;
- possible purposes of a simulation: visualize the law of large numbers, make conjectures, bring out the notion of stochastic model, help solving probability problems...
- the suitability of spreadsheet for introducing continuous distributions (exponential, Gaussian...).

For this, I shall use a theoretical framework including Kuzniak's Mathematical Working Space (MWS), Kuhn's paradigms and Duval's semiotic registers.

Theoretical framework

In order to get a holistic view of the work undertaken by somebody solving a mathematical problem, one has to take into account not only the domains at play but also the cognitive processes involved. The Mathematical Working Spaces, or MWS, framework (Kuzniak, 2011) considers two "planes" –epistemological and cognitive–, each one having three components:

- in the epistemological plane: a set of representations ('real space'), a set of artefacts (instruments) and a theoretical reference system;
- in the cognitive plane, three processes: visualization, construction and proof

An important feature of the model is the interaction between these two planes according three dimensions, semiotic, instrumental and discursive, linking each component of one plane to a corresponding component of the other (Figure 1). The model also assumes that efficient mathematic work results from involving the 3 dimensions together with interactions between them.

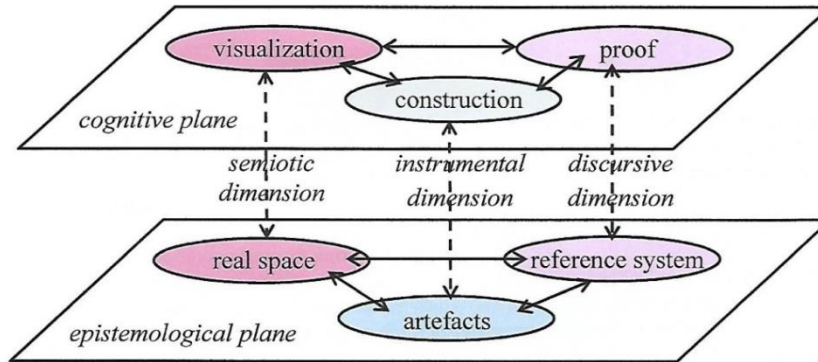


Figure 1: The MWS model (after Kuzniak, 2011)

Kuzniak distinguishes 3 main MWSs:

- reference MWS, defined by the syllabus,
- suitable MWS, planned by the teacher to be implemented in his/her class,
- personal MWS of the student.

Kuhn (Kuhn, 1962) defined scientific paradigms as "universally recognized scientific achievements that, for a time, provide model problems and solutions for a community of researchers," (page X of the 1996 edition). This notion was adapted by Kuzniak to taught mathematics, regarding the epistemological plane. In the case of probability several paradigms can be distinguished (Parzys, 2011):

- a realistic paradigm (R), i.e. the real ('concrete') random experiment itself;
- a paradigm (P1) resulting from a first ("light") modelling of the real experiment by establishing a precise protocol, a list of issues and assigning a probability to each of them;
- a paradigm (P2), in which notions of random experiment and probability are defined, together with properties of probability which can be used for solving problems.
- a paradigm (P3) of the axiomatic type, taught in university.

N.B. In France, at secondary level, only P1 and P2 paradigms are considered, the latter being possibly extended with some elements of calculus at the end of high school (P2+) for the study of continuous distributions.

Regarding the semiotic dimension I shall refer to the notion of 'semiotic register', i.e. a coherent semiotic system allowing 3 cognitive activities: produce identifiable elements (representations), transform an element into another of the same register, convert an element into an element of another register (Duval, 1995). For Duval, a better knowledge is obtained through the use of several registers interacting one with another. He also indicates that the shift from one representation to another one

is more efficient when there is a ‘semantic congruence’ between them, i.e. when there is a one-to-one correspondence between signifying elements of the two representations.

Simulation

In the beginning of its being studied in high school, probability was taught as an application of combinatorics, through Laplace’s formula (probability = number of favorable issues / number of total issues). This ‘cardinalist’ point of view implies that all the issues have the same chance to appear, and then other random phenomena had to be left aside. For instance this is the case for drawing pins: when tossed, they may come down in two ways, like coins, but no argument of symmetry can help and one cannot assign a plausible *a priori* probability to each of them. In such a case you have to observe the relative frequencies of the issues, assuming that they will ‘converge’ toward their probability when the number of tries grows ‘to infinity’. This is the ‘frequentist’ point of view, theorized by the law of large numbers. This point of view was introduced in French high school 20 years ago. Anyway, whatever is the point of view on probability, one has to decide which probability will be allocated to each issue of the experiment, the difference being that the decision is made:

- either on *a priori* ground (e.g. ‘symmetry’ of the issues) in a cardinalist approach;
- or on *a posteriori* ground (frequencies of the issues) in a frequentist approach.

In past days, teachers were reluctant to let their students perform sequences of a random experiment, mostly because it was noisy and requested too much time, but the coming of computers in classrooms, namely spreadsheet including a so-called ‘random’ generator (see for instance Kroese et al., 2011), provided them with an alternative path (although starting with real experimentation remains necessary to materialize the link with reality). Spreadsheet is now widely used to simulate random experiments, with various purposes.

1) Spreadsheet can be used to visualize the compatibility of the cardinalist and frequentist points of view, and finally get the students confident in the generator. For this purpose one has to introduce into the machine a probability for each issue. Then the evolution of the relative frequencies on fairly large numbers of tries can be observed (Figure 2), this process being in fact a visualization of the ‘law’ of large number (belonging to the P3 paradigm).

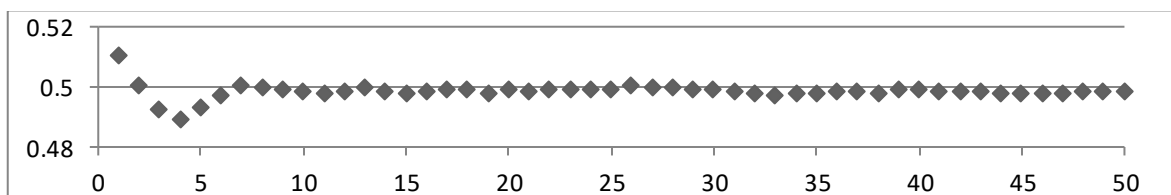


Figure 2: Relative frequencies of heads in 1000, 2000, ...50000 tries of heads and tails

Thus performing a simulation implies constructing at least a simple probabilistic model (within the P1 paradigm), in order to implement it in the machine. Hence simulation is a ternary process: what is implemented in the software is *not* the real experiment but a theoretical *model* of it (Figure 3).

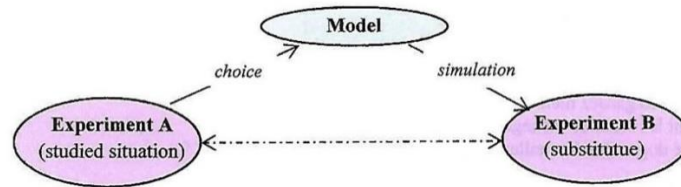


Figure 3: The ternary process of simulation

In such a task several registers are appealed to in turn: natural language, symbolic language (software) and Cartesian graphs. The semiotic-instrumental plane of the MWS is at play, involving the initial P1 paradigm (finding a model of the experiment), then shifting to another paradigm: descriptive statistic (DS) (results of the simulation). It is only when a conjecture about the experiment is asked that the discursive dimension appears (within P2 paradigm).

Both as teacher and teacher trainer, I could observe that some students find it difficult to distinguish between statistical and probabilistic paradigms (the more so than some notions are similar), somewhat analogous with geometrical paradigms. Here, like with GDS, the dynamic feature of spreadsheet, allowing an easy and fast observation of many samples –and consequently many different results (Figure 4)–, can help distinguishing the P2 paradigm (theoretical value) from SD (observed value). But French textbooks do not put the stress on the distinction between the two domains, in particular using the notions in a very loose way (e.g. confusion average / expectation), this probably reflecting actual teaching in classrooms. Similarly, about the instrumental dimension a tendency of textbooks to ‘overguide’ the students, in order to help them deal with the software, must also be noticed.

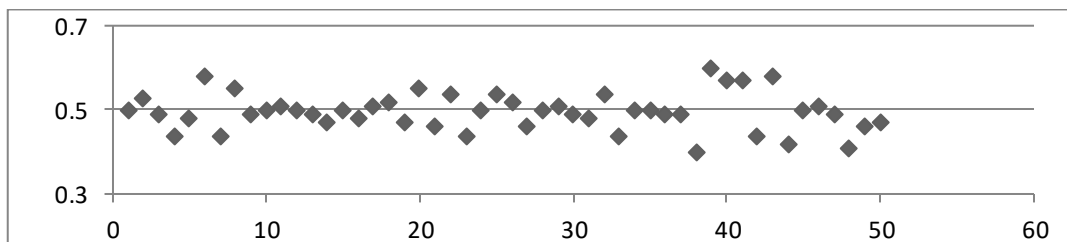


Figure 4: Relative frequencies of heads in 50 samples of 100 tries of heads and tails

2) Simulation can also be used to estimate the value of a probability. For instance, if various models of a same experiment give different probabilities for a given event, a simulation mimicking the concrete experiment can tell which model(s) can be discarded. This is the case with the following problem, which was the basis of an action research with high school students introducing the frequentist approach (Parzysz, 2007):

Toss a well-balanced coin; if head (H) happens you win and if tail (T) happens you toss the coin again; then if H happens you win and if T happens you lose. What is your probability of winning?

In 1754, Jean Le Rond d’Alembert thought that one has 2 chances against 1 to win, at variance with “all authors” claiming that one has 3 chances against 1; Mimicking the process with spreadsheet leads to some difficulty, since one has to distinguish between the two possible results of the first toss. For an easier implementation in the computer the teacher may decide that the coin would be tossed in every case. From a probabilistic point of view the two processes are equivalent but all students are not convinced. (Historically, a similar argument opposed Blaise Pascal with Gilles Roberval in 1654).

The reason for such a reluctance is that the second process is not semantically congruent with the real experiment.

Table 1 shows the corresponding spreadsheets.

Try n°	1st toss	Again?	2nd toss	Result
1	H	no		won
2	T	yes	H	won
3	T	yes	T	lost
4	T	yes	T	lost
5	T	yes	H	won
6	H	no		won
7	H	no		won

Try n°	A coin	B coin	Result
1	H	T	won
2	T	H	won
3	H	T	won
4	T	T	lost
5	H	T	won
6	T	T	lost
7	T	H	won

Table 1: Spreadsheets of the two simulations

When comparing the sheets, one can see that putting anything (H or T) in the empty boxes of the left sheet has no influence on the final result. After that one can forget the “Again?” column and have a second toss in all cases, i.e. replace the initial procedure by the second one without any inconvenience (Figure 5).

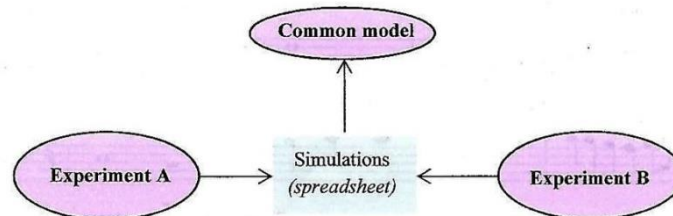


Figure 5: From experiments to model

Thus a visual comparison within the register of double entry tables, in the semiotic-discursive plane, can be a means for deciding if two models are equivalent. And for younger students this can be a possible path towards the notion of stochastic model.

3) A most widespread type of activity in French 10th and 11th grades describes a random experiment and then asks the student to simulate it a number of times with the spreadsheet, observe the results and formulate a conjecture about the probability of one of the issues or the possible value of a parameter. Then he/she is asked to solve the problem using the probability theory and compare the theoretical results with the initial conjecture.

In this process several paradigms are at play. As seen above, starting from reality (R), the student shifts to probability (generally P1), then moves to descriptive statistic (DS) to extract information from the spreadsheet (frequency, mean, etc.) and back to P1 to formulate a conjecture; solving the problem within P2 will imply the discursive dimension (Figure 6).

N.B. In such activities the spreadsheet is used as a multi-purpose tool: it intervenes in turn as logical tool (instructions for software), random generator (simulation), copying machine (results of tries), calculator (statistical parameters) and plotter (diagram).

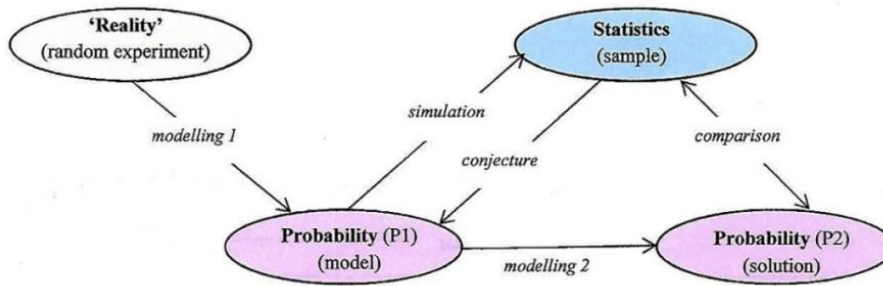


Figure 6: The paradigms and processes involved in the activity

Continuous distributions

N.B. This point is based on a recent research (Derouet & Parzysz, 2016).

The French syllabus for 12th grade includes an introduction to continuous probability laws, and the official resource document for that level suggests starting with a statistical continuous variable, in order to approximate the histogram of a sample by a continuous curve “which fits the histogram, the area under the curve being equal to 1”. The general idea is to link a random variable X , not with a set of isolated probability values as was previously the case, but with a function f (density) verifying : $P(a \leq X \leq b) = \int_a^b f(x)dx$ for any a and b with $a \leq b$. This modelling process, which seems quite sensible, implies also a shift from statistics (DS paradigm) to probability (here P2+) and an essential, though transitional, point is histogram. This rises a difficulty, since spreadsheet *cannot* produce histograms, or rather what it calls histogram is in fact a bar chart. In order to overcome this problem one may widen the bars till they become contiguous (Figure 7), but this trick is restricted to cases in which all intervals have the same width. But the notion of density gets sense only with unequal intervals, since in a histogram the basic notion is area, not height. Thus in this case spreadsheet appears to be of no help if the software does not permit producing *real* histograms.



Figure 7: From bar chart to pseudo histogram

This same syllabus includes the study of normal law and recommends introducing the standard law $\mathcal{U}(0, 1)$ from the observation of the distribution of $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$, where X_n follows the binomial law $\mathcal{B}(n, p)$. Then the bar chart for Z_n is approximated by the curve of a function of the $x \rightarrow \lambda.exp(-ax^2)$ type. As above a shift from discrete to continuous law occurs, but this time spreadsheet can help since, contrary to the general case, the values of Z_n are equidistant (the distance being $1/\sqrt{np(1-p)}$) and then a pseudo histogram is suitable.

All textbooks follow this scheme, in which three types of diagrams are at play in turn: first bar chart (for Z_n), then pseudo histogram, then bell curve. Spreadsheet is necessary at every stage of the process, first to get the values of $P(X_n = k)$ for $0 \leq k \leq n$ and various values of n and p , then the corresponding

bar charts of Z_n , then its pseudo-histogram and finally the (pseudo-)curve of the standard normal law (in fact a polygon). The main difficulty comes from the histogram and the curve looking proportional but *not equal* (Figure 8) because the distance between the values of Z_n is different from 1 (see above).



Figure 8: Pseudo-histogram and pseudo-curve

This problem of scale is tackled rather awkwardly in textbooks, as is the standardization of the binomial variable. However, the question appears when one wants to compare the shapes of the bar charts for several binomial distributions (Figure 9); one may then think of a ‘calibration’, i.e. changing the units on the axes, in order to get diagrams with the same average and height.

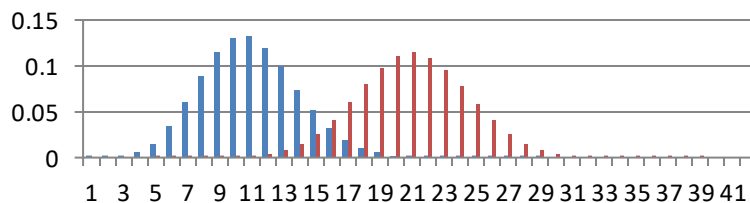


Figure 9: Comparison of binomial distributions ($\mathcal{B}(100, .1)$ and $\mathcal{B}(50, .4)$)

In the process the semiotic and instrumental dimensions of MWS are much appealed to, but the discursive dimension is not much present, due to the students’ lack of knowledge.

Conclusion

The current French high school curriculum starts with descriptive statistics (from 6th grade on) and later goes on with probability (at 9th grade), introduced through a dual, frequentist and cardinalist, point of view involving several mathematical paradigms (DS, P1, P2, P2+). Experimentation has become a central issue in teaching probability and in this process spreadsheet extends real tries, for the reason that it is incomparably faster once its use (language, gestures) is mastered. It is now included in the semiotic-instrumental plane of the MWS and can play an important role in many ways and for multiple purposes. Some points are of importance for teaching with simulation, namely pay attention to the model subjacent to the ‘real’ random experiment (even when it does not clearly appear), help students distinguish between the statistical and probabilistic paradigms, bring out the idea of stochastic model... When coming to continuous probability a sensible way to introduce it consists of approximating a histogram by a continuous curve. Unfortunately usual spreadsheet cannot produce general histograms –i.e. with unequal classes– but only bar charts, becoming possibly ‘pseudo histograms’, and histograms should have to be produced with another software. However class experimentation showed that a pseudo histogram may prove useful as a transitory artefact in the particular case of shifting from binomial to normal law.

On the whole, although spreadsheet was not conceived for educational but for professional purposes, it has now become a quite appreciable, if not indispensable, tool for the teaching of probability in high school.

References

- Batanero, C., Biehler, R., Maxara, C., Engel, J., & Vogel, M. (2005.) Using simulation to bridge teachers' content and pedagogical knowledge in probability. *Proceedings of the 15th ICMI Study conference: The professional education and development of teachers of mathematics*. Retrieved from <http://www.mathunion.org/icmi/digital-library/icmi-study-conferences/icmi-study-15-conference/>
- Carranza, P. & Kuzniak, A. (2008). Duality of probability and statistics teaching in French education. 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. Monterrey: ICMI and IASE. Retrieved from http://iase-web.org/documents/papers/rt2008/T1P2_Carranza.pdf
- Derouet, C., & Parzysz, B. (2016). How can histograms be useful for introducing continuous probability distributions. *Zentralblatt für Didaktik der Mathematik*, 48(6), 757–773. Dordrecht: Springer.
- Duval, R. (1995). *Sémiosis et pensée humaine*. Bern: Peter Lang.
- Kroese, D., Taimre T., & Botev, Z. (2011). *Handbook of Monte Carlo methods*. New York: Wiley & Sons.
- Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago.
- Kuzniak, A. (2011). L'espace de Travail Mathématique et ses genèses. *Annales de Didactique et de Sciences Cognitives*, 16, 9–24.
- Parzysz, B. (2007). Expérience aléatoire et simulation: Le jeu de croix ou pile. Relecture d'une expérimentation déjà un peu ancienne. *Repères-IREM*, 66, 27–44.
- Parzysz, B. (2011). Quelques questions didactiques de la statistique et des probabilités. Strasbourg, *Annales de Didactique et de Sciences cognitives*, 16, 127–147.
- Pratt, D., Davies, N., & Connor D. (2008). The role of technology in teaching and learning statistics. In Batanero, C., Burrill, G., Reading, C. & Rossman, A. (Eds.). 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. Monterrey: ICMI and IASE