"Science Stays True Here"
Journal of Mathematics and Statistical Science, 210-220 | Science Signpost Publishing

Using Simulation by Excel and R in the Teaching of Probability and Statistics for Non-Math Major Students

Bui Anh Kiet

Cantho University

Jean-baptiste Lagrange

LDAR, Université Paris-Diderot, France

Abstract

In Vietnam, the teaching of probability gives an important place to a «classical» approach based on the formula of Laplace, and inferential statistics is seen as an application. Moreover, the teaching does not include the use of software, generic or dedicated to statistics. This leads to difficulties that have special meaning for students of other disciplines as mathematics, to be prepared for an understanding of random phenomena and a statistical approach in real life. This article aims to assess the possibility of introducing "viable" innovations in teaching probability and statistics at the university level in Vietnam, and the improvements that these innovations bring to the understanding of the students.

Keywords: *frequentist approach, probability and statistics, R, spreadsheet, models*.

Introduction

In Vietnam, as in other countries, textbooks and the curricula in probability and statistics at the university are developed in this order: combinatory analysis, the formula of Laplace, the relative frequencies in repeated trials, calculation on events, random variables, then descriptive and inferential statistics.

Textbooks and curricula focus on a classical approach, that is to say the application of combinatorial analysis using the formula of Laplace in situations where outcomes are equally likely. In the chapter on the relative frequencies, they admit that the probability of an event can be obtained as a “limit” of the relative

Corresponding author: Bui Anh Kiet, Cantho University.

frequency of this event in repeated trials, and provide a limited number of "classical" examples such as tossing of a coin, but do not offer any tasks in this chapter. Descriptive statistics are considered independently of probability and inferential statistics are presented as an application of probability theory.

This causes difficulties for students:

- Even for "classical" tasks, many students have difficulties with combinatory analysis and therefore cannot successfully calculate theoretical probabilities. They have no means to confront their reasoning and calculations to a reality, since no connection is made with empirical relative frequencies.
- Students do not integrate the connection between the empirical relative frequencies and the theoretical probability of an event.
- Students see inferential statistics as an application of probability theory. However, it is a very difficult topic at a theoretical level, and in most cases, they are only able to use inferential statistics at a practical level, that to say applying formulae without understanding the underlying theories.
- In addition, students are not really confronted to random phenomena, like the fluctuation of relative frequencies and the relationship between the fluctuation and the size of a sample. It means that opportunities are missed for a deep understanding of probability theory and its connections with statistical questions arising in everyday live, like the size of a sample to get an estimation of a probability at a given accuracy and at a given confidence level.
- Textbooks and curricula do not take into account the growing use of software, generic or dedicated to statistics, and thus appear to students far from practices in the professional world.

These difficulties have special significance for non-math major students, who must be well prepared for the understanding of random phenomena and a statistical method in the real world and the professional world, rather than a purely mathematical approach.

The overall objective of the study is to overcome these difficulties and, specifically, to evaluate the possibility of implementing two innovations, one linked to a "frequentist approach", and the other to the simulation, within an existing course for non-math major students.

Literature Review and Objectives of the Study

There is strong research evidence that misconceptions about probability do not disappear as a result of traditional education focused on formal definitions, rules and procedures. Although students can learn the rules and procedures in probability and even if they get accurate answers to math tests, these same students often misunderstand about ideas and basic concepts and often ignore the rules when they own

judgment on uncertain events. Especially for university students, misconceptions can also appear as the result of education received in the form of explicit theories used outside their scope. A pedagogical intervention specifically designed to eliminate misconceptions students on probability is necessary for tangible and stable improvement in the concepts students can be obtained. The active participation of students in the construction of knowledge, the confrontation with large samples, and the use of computer simulation are tools for that.

We also retain the idea that the relative frequency of an event and the long-term patterns of behavior play a very important role. We want to introduce a “frequentist approach” of the probability of an event based on observation of the convergence of relative frequencies for this event in repeated randomized trials in order to sensitize students on both unpredictable in the short term random phenomena and the long-term regularity that describes the probability. In addition, the frequentist approach will be helpful in providing an approximation of a real probability, with a large enough sample: students will be able to compare results obtained by empirical observation frequencies and the theoretical results obtained by the classical approach. Observing a divergence can help them become aware of a misconception.

The use of simulation is consistent with the frequentist approach, because the observation of convergence relative frequencies is generally not possible on the actual data. Furthermore, simulation is used to build a model of a random phenomenon, giving a deep understanding of the underlying random situation, and preparing for a “classical approach”. According to the literature, such as Garfield, Chance & Snell (2000), a spreadsheet like Excel, and a programming language like R provide opportunities for this. With a spreadsheet students can simply build a simulation and directly view the results. They can use the F9 key to create new samples and observe the changes in relative frequencies between samples. R is a professional statistical software and therefore its use helps students prepare for new uses in their professional lives. In addition, as R is a structured language, students can use a modular approach in the construction of a model. This helps them to understand the deeper conceptual model. In contrast with a spreadsheet, simulation in R may be performed on very large samples.

The research questions arise from these considerations:

- What are tasks and techniques related to a "frequentist approach" and how they improve the teaching / learning of probability, particularly regarding misconceptions and inadequate models random situations?

- How to connect the frequentist approach and the classical approach? Especially, how to build a milieu (Brousseau, 1997) to implement adequate and appropriate didactical contracts in this connection?

Methods

Our study is exploratory, in the sense that we have implemented experimental sessions designed and conducted by the first author of this article as part of his dissertation, and that these sessions serve both to highlight the contributions of innovations implemented, and identify missed opportunities that it is analyzing as points of attention for future implementations.

The Experimental Sessions

The sessions were conducted at the University of CanTho in a class of 30 students of disciplines: economics, engineering, computer science, and agriculture ... These students follow the course of probability and statistics (the first year) organized a series of 15 lessons (45 periods, 50 minutes/period) for 15 weeks. Three lessons were consolidated for four weeks to organize four experimental sessions of two and a half hours. The four experimental sessions correspond to four random problems. In each problem, we consider one or more events, or a random variable and ask a question related to the probability of these events, or the expectation of the random variable.

We choose these problems because in each of them, the probability of event(s) is not obvious to students and may be a subject of a debate. For each problem, the construction of a simulation must provide two different contributions: first to get the relative frequencies or averages of an event for a given number of trials, and then to develop a model that is also useful for a theoretical calculation of the probability. The joint implementation of the four sessions following this analysis: students first do some practical experiments to become familiar with the problem then the teacher starts a discussion on the problem, after that the students build a simulation and observe relative frequencies of an event or empirical average of a random variable with questions of fluctuations. The last step is the classical mathematical calculation to confirm the answer.

The Problems

1. The two dice problem

Let X be the sum of the numbers appeared after rolling of two fair dice. The question is to compare

P(X = 7) and P(X = 8). These numbers were chosen because a common misconception is that these probabilities are equal, the pair (4; 4) being counted twice in the numbering of elementary events.

2. The Rabbit and Turtle problem

Rolling a fair die, if the number 6 appears, the rabbit wins; if the number 6 is not displayed then the turtle stepped. Continuing to roll the die until the rabbit wins or until the turtle gets six steps and wins. The question is to decide who is more likely to win. The number six was chosen such that the answer is not obvious.

3. The hunting duck problem (Engel, 1990)

There are five hunters and five ducks. Each hunter shoots one of five ducks at random and never misses it. The question is the average number of surviving ducks. The number 5 was chosen so that the size of the sample space is large enough (5^5 = 3125) that simply counting is unrealistic.

4. The Monty Hall problem

There are three doors, behind a door is a car and behind the other two doors is a goat. The player is allowed to choose one of three doors and it is not open. Then the host opens one of the other two doors. In the case where the car is behind one of the two non-selected doors, the host always opens the door without the car. Then the player can remain on his first choice to select or change the other door is not opened. The question is whether, to win the car, the player must keep his original choice or switch to the other door.

Simulation and Software

To introduce the use of software in the experimental sciences, we use the concept of the didactical functionality of a digital tool (Cerulli *et al*., 2006) defined by three key elements.

(1) A set of features / characteristics of the tool

Like all spreadsheets Excel is software based on formulae written in a table. We can see the evaluation of the formulae dynamically updated on the screen of the computer. Repetition is done by dynamic copy of formulae in liens or columns and is therefore limited in practice. It does not require writing a program as programming language like R. Excel also has standard functions for descriptive statistics (mean, countif) and a pseudo random generator under two forms (RAND, uniform distribution of decimals between 0 and one; RANDBETWEEN(a,b); uniform distribution of integers from a to b). The graphic display of Excel is good for presenting data collected from simulations. Excel also has the function key F9 that is used to recalculate all formulae in a worksheet, including new values for the data issued of pseudo random generation. R is a functional programming language. R can be used by command

lines or by programming structured user functions. R has an alternative structure (if ... else…) The main structure for repetition is the for loop. R has functions for vectors (mean, unique) useful for statistical treatments. The pseudo random generator is called by way of a function sample(a:b, n, repl). If a and b are integers, the function returns a pseudo random vector of n uniformly distributed integers between a and b. The parameter repl controls the "replacement". Resampling can be performed by reexecuting.

(2) An educational aim

According to the law of large numbers, Excel and R can be used to simulate a random situation and collect data in order to reconcile the theoretical probability of an event or the theoretical mean of a random variable. The simulation results in Excel or R will be used to confirm or deny the theoretical calculation of probability and help students better control these calculations. This is the pragmatic dimension. The simulation in the context of education can also have a conceptual dimension. First, students can become aware of the process of relative frequencies or statistical means tending to limits, when the sample size increases and the fluctuation decreases, understanding these limits as probabilities or theoretical expectation. Second, to build a simulation, students will need to build a model. We expect that reflecting on models will help student a better theoretical control of standard calculations in probabilities.

(3) Modalities of employing the tool in the teaching / learning process.

In each experimental session, students will first make a simulation with Excel and a simulation with R, before making a theoretical calculation. They will thus build a first spreadsheet model and get results on limited samples, and then adapt the model or create a new one with R and get results on larger samples. The models will be useful for the theoretical calculation and simulation results will be compared with values obtained by theoretical calculation.

Experimentation

For each session, we give some indications on the course, and then a summary analysis of the observations made by Bui (2015).

Experimental session 1: The sum of two dice

The teacher helps students become familiar with the realization of a concrete experimentation with two dice and a data collection. He uses collected data so that students see the need for a larger sample size in order to compare the odds of getting respectively sums of 7 and 8. Then students perform simulations using the Excel spreadsheet guided by the teacher. Fluctuating of relative frequencies leads students to see

the need for a larger sample size. Then, the teacher presents the R software for the simulation. At the end of the session, he discussed with the students about the value of the simulation.

Students are sensitive to fluctuations and link with sample sizes but without exploiting them directly to compare the two involved probabilities. Regarding the sample size in the simulations, the teacher and students are limited to two sizes: 1000 (too small) and 100,000 (big enough); the opportunity has been missed to question the sample size in order to determine an optimal size for discriminating the two probabilities into the problem play ($P(S = 7) > P(S = 8)$ in 95% of cases).

Students have difficulties to draw random numbers in the spreadsheet but they do not encounter many difficulties when programming with R.

This experimental session reflects the misconceptions of students in different phases. These misconceptions are destabilized by the confrontation with the relative frequencies obtained by simulation on a large sample, rather than a reflection on the model developed for this simulation. In particular, in the last phase when calculating theoretical probabilities, the relative frequencies obtained by the simulation are used to invalidate the misconception of the sample space consisting of pairs (unordered). However, the model developed for the simulation is not used, when considering for example the data generated by the spreadsheet, students could see that the sums 4 + 3, 3 + 4 and 4 + 4 appear with the same relative frequency. The teacher institutionalizes the pragmatic role of simulation rather than a reflection on the model built in phases of simulation.

Experimental session 2: The rabbit and the turtle

This session marks a progression in complexity and is not easy for many students. After the first phase of experimentation with a die, the teacher and students do not show the original question (the most likely winner) that is thus forgotten in favor of a centering on the probability of each event (the rabbit wins, the turtle wins), first approached frequentist way, then theoretically calculated. This is the result of a didactical contract oriented probability values and minimizing the statistical questions.

As in the previous session, students have difficulties with the spreadsheet. The model implemented with the spreadsheet, conforms to the situation in the sense that the process is stopped after a win, which is not easily realized in the organization of the spreadsheet. This stop condition is not convenient to implement in R and another model is proposed which is to roll the die 6 times, and conclude that the rabbit wins if there is a 6 among the numbers obtained. This model is again proposed by the teacher for the theoretical calculation to guide students in the use of the formula of Laplace, but it is quickly abandoned after a student proposes to use the probabilities multiplication formula. Thus, different models are used in

different phases. This change of model simplifies programming in R and could simplify the theoretical calculation, but it is not the subject of a class discussion: the new model is not compared to the previous model and the equivalence of models is not discussed. This is again a missed opportunity, and as in the previous experimental session, the pragmatic role of simulation is favored by the teacher at the expense of a reflection on the models used in the simulation phase.

Experimental session 3: the hunting duck problem

Students do not encounter many difficulties in the use of Excel and R functions to simulate. As in previous sessions, the teacher stresses the need to increase the size of the sample appropriate to the simulation for the transition to the use of R. In this way, students offer a very large sample size without there is a discussion of an appropriate size. This analysis confirms that the simulation can provide a rich environment for student reflection. However, it seems that the teacher can do better by asking the minimum sample size for a given precision.

Strategies for calculating the average are different in the two simulations: with the spreadsheet, the average is calculated directly by globalizing draws (total number of ducks killed divided by the number of shots); with R, the students first calculate the empirical relative frequencies of each of events, after the presentation of the mean function by the teacher, the students drop relative frequencies and directly calculate the mean as the spreadsheet. For the theoretical calculation, the teacher leads students to calculate the probability distribution of number of ducks killed by comparing the empirical relative frequencies obtained from R as a prerequisite to the calculation of the expectation. The failure to discuss the possibility of direct calculation of expectation bypassing the probability distribution is another missed opportunity: in fact, the discussion on this topic could focus students' attention on the expectation in repeated draws (sum of expectations in each draw).

Experimental session 4: The Monty Hall problem

From experiments without software, the students understand the situation and recognize that the choice to change door will bring more chance of winning. The simulations that follow with larger sample sizes to help students eliminate erroneous intuitions and quantify more precisely the winning probabilities for each choice (keep the same door or change). In each case, the simulations were made for each of the two choices. Implicitly, the choice of the door opened by the host if the player chose the winning door is equally likely.

The first simulation with Excel leads to quite complex formulas, but students are able to present many different solutions. One student remarked that if the player retains the same door he wins if and only

if he has chosen the winning door in the first choice. This significantly simplifies the simulation and is used by most students in the simulation with R.

However, for the theoretical calculation, the teacher engages students to a calculation using conditional probabilities conditioned by the choice of the host. These conditional probabilities are calculated by applying academic knowledge already taught in the course but do not respond in fact to the same question asked and lead to the same value only because equi-probability is assumed for the choice of the host (Grinstead and Snell, 2005, p.139). An opportunity to discuss models, linking simulation and theoretical probabilities was also missed in this situation.

Discussion and Perspectives

Tasks and Techniques

The simulation brings new tasks associated with statistical questions. The choice was made in experimental sessions to offer these tasks before classical tasks. We observed that students routinely check their results against the empirical results, particularly in the third session where five theoretical probabilities are calculated, and also to destabilize misconceptions or false models. This check appears as an integrated technique in action programs. This means that students articulate simulation with classical techniques already taught, to obtain better control of these techniques. The practice of this «blended technique» seems to have a positive effect in respect of misconceptions: observation shows that students question their models from data of simulations. It also has limits: a student can adapt wrongly a wrong model in order to get a theoretical value consistent with data obtained by simulation.

Milieu and Didactical Contracts

In experimental sessions, the simulation can be regarded as a milieu at two levels:

1. A milieu for action

The relative frequencies and their fluctuations can be considered that students get feedback during the construction and execution of simulations on the computer, which reinforce their understanding of the relationship between relative frequencies and probabilities. Implicitly, the simulation allows students to build models "in action" that destabilizes misconceptions. The simulation also plays a role as the milieu of the action in the calculation of theoretical probabilities when students systematically check their results against empirical relative frequencies.

2. A milieu for reflection

We saw that during each session, several models of the random situations were considered, but they are not discussed. Moreover, we have seen that the fluctuation is not studied precisely as regard to questions initially asked: the teacher promotes in fact a practical dimension of the simulation, emphasizing an appropriate size of sample to motivate the use of R, rather than to discuss relevant size with respect to the statistical question. This shows that potentially, the simulation helps to create a milieu of reflection that has not been exploited: the existence of several models should be an opportunity to discuss their equivalence; discuss sample size in relation to the posed question should be a preparation for inferential statistics. We have identified several "missed opportunities" that can be interpreted as an underestimation of this "reflection" dimension of the milieu.

The questions posed in the 4 experimental sessions are problematic for students and motivate building simulations. However we note that in these sessions, as soon as they pass the simulation and calculations, students and teachers no longer refer to the original question. Thus, these initial questions generate a motivation so that students find probabilistic values, rather than a true statistical survey. We interpret this as the manifestation of a didactical contract oriented towards practical use of simulation to approximate probabilistic values rather than to investigate the issue in question, under the influence of a "traditional" didactical contract promoting the calculation of theoretical probabilities.

This critical analysis allows us to offer a reconstruction from all 4 experimental sessions in order to better exploit their potential by putting more emphasis on probabilistic issues and taking into account further simulation as a milieu of reflection on models of probabilistic situations.

References

Batanero, C., & Sanchez, E. (2005). What is the nature of high school students' conceptions and misconceptions about probability? In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 241–266). New York: Springer.

Bui, A. K. (2015). Apports de la simulation et de l'utilisation de logiciels pour l'enseignement /apprentissage des probabilités et des statistiques en première année d'Université au Vietnam dans un cursus non mathématique. *Thèse de doctorat*. Université Paris-Diderot.

Brousseau, G. (1997). *Theory of didactical situations in mathematics*. Dordrecht, The Netherlands: Kluwer.

Cerulli, M., Pedemonte, B., & Robotti, E. (2006). An integrated perspective to approach technology in mathematics education, In Proceedings of CERME 4, Sant Feliu de Guíxols, Spain (pp. 1389-1399).

Engel, A. (1990). *Les certitudes du hasard* (Certainty of chance). Aléas Editeur, Lyon.

Garfield, J., Chance, B. L., & Snell, J. L. (2000). Technology in college statistics courses. In D. Holton et al. (Eds.), *The teaching and learning of mathematics at university level: An ICMI study* (pp. 357–370). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Grinstead, C. M. & Snell, L. J. (2005). *Introduction to Probability* (2nd ed.). New York: Random House.

Parzysz, B. (2009). Des expériences au modèle, via la simulation (From experiences to a model, via simulation). *Repères-IREM, 74*, 91–103.

Published: Volume 2017, Issue 7 / July 25, 2017