Mathematical modelling in probability at the secondary-tertiary transition, example of biological sciences students at university

Camille Doukhan

Mathematical modelling in probability at the secondary-tertiary transition, example of biological sciences students at university

Camille Doukhan[1]

[1]CREAD, Université de Brest, France, camille.doukhan1@inspe-bretagne.fr

Abstract. The work presented here focuses on probabilistic modelling at the secondary-tertiary transition. Concerning the university level, I am interested in non-specialist students, more precisely first-year biology students. I have written and submitted a test on discrete probabilities to Grade 12 and to biology students. I present here the a priori and a posteriori analysis of one exercise of this test. The analysis of the students' responses to this exercise allowed me to establish initial results on possible difficulties in the secondary-tertiary transition. A first point is that students, in these modelling activities, have difficulty linking task and type of task. Another result is the very high use of probability trees by students in their modelling of the probabilistic situations proposed.

Keywords: mathematics in other disciplines, modelling, probabilities, secondary-tertiary transition, students' activity

INTRODUCTION

The study of the secondary-tertiary transition is not new in mathematics education research (Gueudet & Thomas, 2019). Many authors have focused on the difficulties encountered by students in the fields of calculus or linear algebra (e.g. Vleeschouwer & Gueudet, 2011). There are fewer studies on the topic of probability. However, probabilities are taught in many university courses, particularly for non-specialists such as biology students. My study takes place in France and I have chosen to focus on biology students who study probability from the first year of university.

In the following section, I present the context of my research and previous works about the teaching and learning of probabilities on which I based my study. Then in the third section I present my theoretical framework. In the fourth section I present the methodology I used in my research. In the fifth section I present my results. Finally, in the last section I present the conclusions of this study.

RELATED WORKS

I focus here on the secondary-tertiary transition in the particular context of mathematics in service courses, especially for biology students.

The relationships between biology and mathematics have been studied extensively (Lange, 2000) because they are quite complex and important. Biology has long been a mainly descriptive science and the recent development of new mathematical modelling tools has had an impact on this science by making it highly mathematical, especially giving to it a predictive and decision-making role. Mathematical modelling and probabilities are widely used by biologists (Duran and Marshall, 2019).

The good mastery by biology students of these mathematical contents is an objective of the biology studies at university. That's why general probability courses are designed for biology students from the early years of university to allow them to take courses in statistics or biostatistics in subsequent years.

Several issues are raised by the teaching of mathematics at university for non-specialist students. The difficulties encountered by non-specialist students in mathematics courses would be an important factor in dropping out of their programmes. Previous works have evidenced that the lack of links between these mathematics courses and the future professional practices of these non-specialist students. This lack of links leads them to consider mathematics as too abstract, and to have difficulties to mobilize mathematical tools (González-Martin, Gueudet, Barquero & Romo-Vazquez, to appear).

Concerning more precisely biology students, Viirman and Nardi (2018) highlighted that their involvement in mathematical modelling activities is a motivating factor in their learning of general mathematics courses.

I have therefore chosen to focus my work on mathematical modelling aspects and in particular the use of probability trees. In order to expose my analyses in the fourth part, I will present in the following my theoretical framework.

THEORETICAL FRAME

In this research, I choose an institutional perspective and consider that secondary school and university are two different institutions.

I use the Anthropological Theory of Didactics (ATD, Chevallard, 2006) and more particularly the concept of praxeology. A praxeology consists of four elements: a type of task; a technique to accomplish this type of task; a technology which is a discourse explaining and justifying the technique; and a theory. The comparison of praxeologies in secondary school and at university is very useful because it allows me to highlight possible difficulties for students during this transition. For example, it is interesting to consider the technique produced by the student when it is not explicitly requested; or to look at the technology used- or not- by the student in his/her solution of an exercise.

I also use Activity Theory and its adaptation to mathematics education (Vandebrouck, 2008) to allow me to look more closely at the complexity in the student's activity for a given task. I use here the notion of task as described in the Theory of Activity, i.e. referring to the object of the activity and its description. This theory has allowed me to highlight that the complexity, for a student, of linking a task proposed to a type of task is a feature of the secondary-tertiary transition. In my analyses below, I will give examples of this process.

MATHEMATICAL MODELLING IN PROBABILITY

I consider here the activity of mathematical modelling in general for the theme of probabilities (it can be the activity expected by the text of an exercise or the actual activity of students). This is what I call "mathematical modelling in probability".

Usually it starts with a random situation described in natural language. It is then necessary to identify the events at stake, name them and determine their probabilities. According to the situation, the use of a probability tree can be relevant or not. Mathematical probability modelling mixes recognition activity (recognizing the task to be accomplished and linking it to a certain type of task for which a technique is known), changes of register (moving from natural language to probability formalism) and entanglements of techniques (I will illustrate this probabilistic modelling activity later on).

I claim that associating concepts from the Activity Theory and the Anthropological Theory of Didactics contributes to a precise understanding of this mathematical modelling in probabilities and will illustrate it below.

Based on the theoretical tools developed above, I present as follows my research questions:

How do students build a probabilistic model for a situation from a biological context?

How to describe these probabilistic modelling activities using praxeologies?

METHODOLOGY OF THE RESEARCH

My research takes place in France, in secondary school in a rural environment and in a middle-sized university. Mathematics courses are offered to biology students from the first year, including a course entirely devoted to probabilities (14 hours of lectures with about 300 students and 14 hours of tutorials, which are sessions dedicated to exercises in groups of 30 students).

For this research, I was particularly interested in this course and attended two lectures and two tutorial sessions on the following probability topics: "independence and conditioning" and "continuous random variables". I also observed a class of Grade 12 in the science section of the secondary school (called "scientific section") for seven one-hour sessions on these same probabilities' themes.

I chose these two probability themes because they are part of the Grade 12 curriculum and are taught again at the university for biology students. Probabilities are present in the secondary school curriculum (there are discrete random variables in the Grade 11 curriculum, conditioning and independence in Grade 12 and continuous random variables in Grade 12 as well) (Ministère de l'Education Nationale, 2011). The probability course in biology at the university is quite general, it includes notions seen in secondary school (conditional probabilities and independence, discrete random variables, continuous random variables) but it also contains chapters devoted to new concepts (such as probabilistic model and probabilized space, independence of random variables, limit theorems and their applications, law approximation).

I have chosen to study the secondary-tertiary transition through the chapter of discrete probabilities "independence and conditioning" common to both curricula. In a previous study (Doukhan & Gueudet, 2019), I have evidenced through textbooks analyses that the same probability content (discrete random variables in this study) can lead to very different praxeologies in secondary school and at university. Applying similar methods here (analysis of the resources collected during my observations, such as films of the course, course handouts and textbooks) evidenced the interest of a focus on modelling and probability trees.

I designed a test and submitted it to secondary school students and to first-year biology students. This test (see Appendix) consists of three exercises on the theme "Independence and conditioning".

The test was administered to all the students (29) from the class of Grade 12 observed, during the fifth session on the topic of "Independence and conditioning", students had half an hour to do the test. The test was also offered to first-year biology students prior to their probabilities course of the second semester, with 25 of them participating for a similar duration.

I carried out a quantitative analysis of all the students' productions, which allowed me to highlight first results concerning mathematical modelling in probability at the secondary-tertiary transition, I also analysed students' answers. I have chosen here to focus only on the analysis of the second exercise, which reads as follows: "You are the director of the Minister's Office of Health. A disease is present in the population, in the proportion of one sick person out of 10,000. The manager of a major pharmaceutical company comes to you to tell you about his new screening test: if a person is sick, the test is positive at 99%; if a person is not sick, the test is negative at 98%. Do you authorize the marketing of this test?". This exercise is a classic application of the Bayes' Theorem with an important modelling work left to the student: statement is in natural language, events to be identified, etc. (more details in the a priori analysis in the following section)

I have chosen this exercise because it requires an important probabilistic modelling work, and then because is linked with a biology context.

MAIN RESULTS

A priori analysis

I present here the *a priori* analysis of the second exercise of the test.

The main types of tasks I have identified for this exercise are, first, to perform probabilistic modelling based on a natural language statement; second, to calculate a conditional probability; and finally, to interpret the numerical result in order to answer the question in natural language.

Each of these types of tasks contains several subtypes of tasks with which particular techniques are associated. The difficulty of each subtype of task depends on the precise teaching context. In order to analyse how students will link tasks and types of tasks, I have to associate AT and ATD because an analysis in terms of ATD would be insufficient for my purpose, for this reason I also use task analyses developed in Activity Theory.

The associated subtypes of tasks for the first type of task: "perform probabilistic modeling of a natural language statement", are as follows: identify probabilistic events, associate their probabilities to each of the events, identify contrary events and calculate their probabilities. All these types of tasks entirely within the scope of a probabilistic modelling activities as defined in a previous section.

For the second type of task "calculate the probability of being sick knowing that the test is positive", the associated subtypes of tasks are as follows: interpret the question in terms of probability, calculate the corresponding conditional probability.

The task is complex, as there are many subtypes of tasks for each of the task types identified above. In this exercise, identifying all these subtypes of tasks and organizing their reasoning are the student's responsibility. Here the combination of AT and ATD allows me to see how the complexity of the task impacts the praxeological organization and in particular the complexity of the technique to be implemented.

The techniques associated with these types of tasks are as follows: identify the events involved, associate the numerical data of the statement with events, calculate the missing probabilities (use of the probability of the complementary), identify the probability that must be calculated in order to respond, calculate this probability (for this it is necessary to calculate an intersection and use the Bayes formula); finally, interpret the result. The technique of representing the situation by a probability tree is a technique expected in secondary school but is no longer part of the praxeology at university. Here is a representation of the situation by a tree:

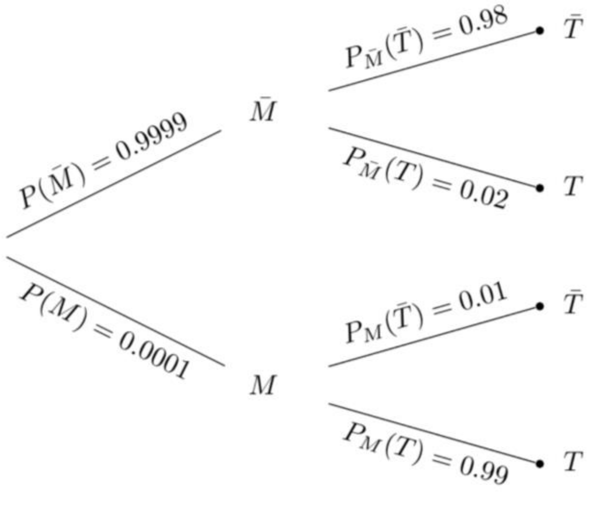

Table 1: probability tree

Here are some technological elements that justify the choice of the probability to be calculated, $P_T(M)$. What I am interested in here is whether the test is effective from the point of view of a caregiver. A patient comes for a test to find out if he/ she is sick or not. If his/her test is positive (respectively negative), it is important to know if he is really sick (respectively not sick), i.e. if the test is reliable. The test is reliable if the probability, knowing that the test is positive, that a person is indeed sick, is close to 1. The goal is therefore to calculate $P_T(M)$, for this it is necessary to calculate an intersection and use the Bayes formula, their uses can be justified by the associated theory or by the use of a probability tree. These technological elements go beyond mathematics, the fact of calculating $P_T(M)$ is entirely at the students' expense and is based on technological elements linked with a socio-medical context.

The context of the exercise, which is the study of the effectiveness of tests on sick and non-ill populations, is rather a context that is familiar to students. Indeed, there are many exercises in this context in secondary school textbooks. In this exercise they have to interpret the question by proposing the probability of an event themselves, then interpreting the numerical result obtained. In secondary school textbooks it is rather common to have in the first question "calculate the probability of such an event" and in the second question "interpret the result". Here, therefore, calculating $P_T(M)$ requires an important initiative from the student.

The techniques to be used by the students are all based on knowledge being acquired and already applied in other situations encountered in secondary school. On the other hand, recognizing this complex praxeological structure and organizing oneself accordingly is entirely at the student's expense because there is no intermediate subquestion associated with each of the subtypes of tasks described above. The very strong modelling activity left to the student is not something usual for them, so I expect to find difficulties for this exercise in the *a posteriori* analysis.

A posteriori analysis

I present here the *a posteriori* analysis of the second exercise of the test.

This exercise was tackled by 91% (49 students) of the students who answered the test (54 students). Both Grade 12 students and biology students encountered a lot of difficulties and proposed erroneous solutions. I expected these results from my a priori analysis above. The diversity of responses is very high, the probability most often calculated by Grade 12 students is P(M and T) (5 responses); there is no dominant answer among biology students. An example is presented in Figure 2.

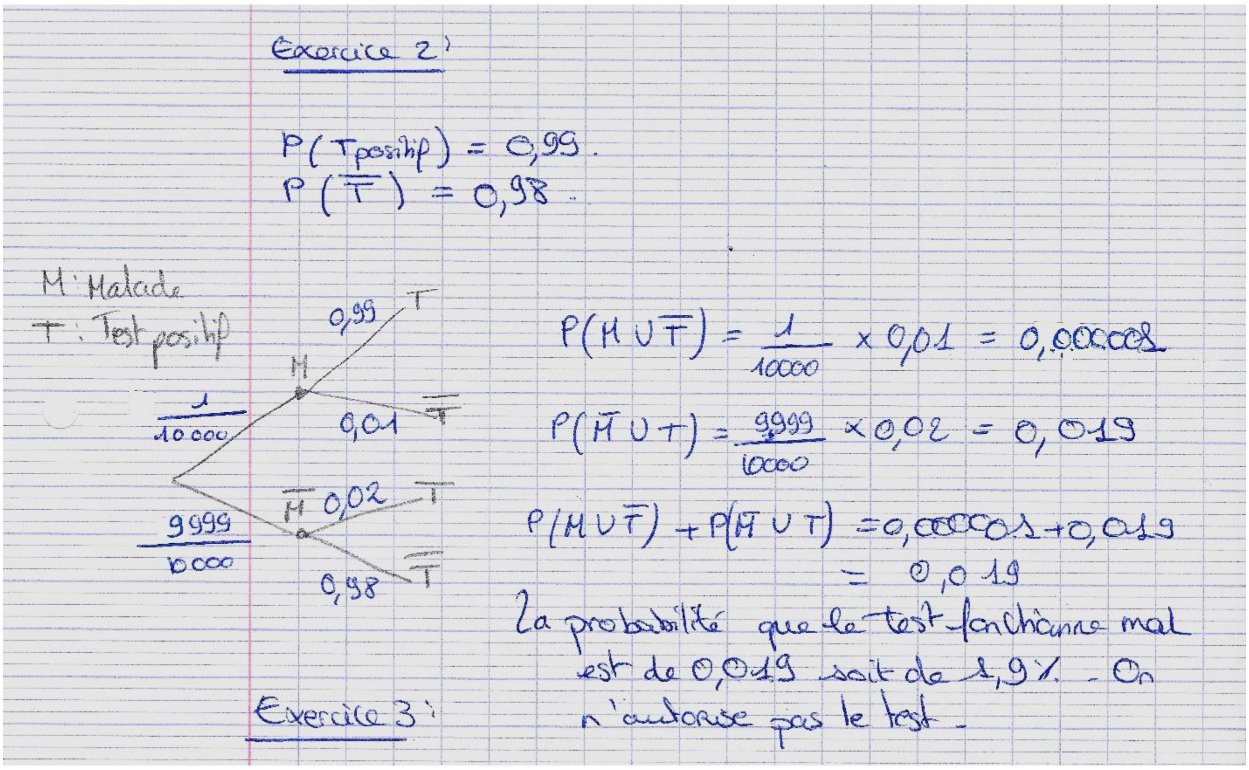

Figure 2: Example of a student's production. "The probability that the test goes wrong is of 0.019, meaning 19%. We do not authorize the test"

Even if the students did not answer correctly many of them have correctly identified the events at stake (32 of them, or about 65%). A slightly fewer amount of them represent the situation with a probability tree (26 of them, or about 53%). On the other hand, all the students who chose to represent the situation by a tree correctly identified the numerical values of the statement with the corresponding events and correctly calculated the probabilities of the complementary events. Considering here that correctly modelling the situation means identifying all the events at stake and associating their probabilities to them; therefore, in this exercise, only half of the students who responded did a correct modelling of the situation.

Concerning the interpretation of the question and the answer given by the students, 41 of them (or about 84%) have formulated a response in natural language. Of these, 17 relied on their previous probability calculations to answer. In contrast, 8 students answered the question in natural language based solely on their representation of the situation through a probability tree. Sixteen of them (about 33%), 10 Grade 12 students and 6 biology students answered the question without having previously made any probability calculations or probabilistic modelling (like probability tree). These students were unable to identify in the task prescribed to them the different subtypes of tasks to be performed. Here is an example of such a response:

> "No, because the margin of error is enormous for a population of 10,000. Out of 10000 there could be 200 people who are reported as sick when not at all. This is related to the 98%. If the disease was 1 in 100 people, it would have been more interesting."

More specifically, with regard to probabilistic modelling activities, students highly use probability trees in this exercise while there is no indication anywhere in the statement that a tree could be used to answer as can be seen on the student copy excerpt (see "table2"). This is an important observation which is linked with a secondary-tertiary transition issue.

Indeed, since secondary school, students have become accustomed to use this type of representation. The construction and use of a probability tree are skills that are widely developed in the official mathematics curriculum of the Grade 12 class (Ministère de l'Education Nationale, 2011).

Probability trees also play a very important role in the Grade 12 course I observed, the outline of the course handout distributed by the teacher for the chapter "Conditional Probabilities", consists of three main parts: "Conditional Probabilities", "Probability Trees" and "Independence of Two Events". The rules for building the tree are detailed with technological elements, here is an example: "Rule 3 (total probability formula): the probability of an event is equal to the sum of the probabilities of each of the paths leading to it". Moreover, the official Grade 12 mathematics curriculum states that: "a properly constructed probability tree is a proof", which is no longer conceivable at the university.

CONCLUSION

I have seen through the analysis of students' answers of this exercise that probabilistic modelling is an important issue in the secondary-tertiary transition.

First, it should be noted that students seem to have appropriated the use of probability trees. Indeed, through the analysis of this exercise and the two other exercises of the test that I have not developed here, I saw that the use of trees allowed students to respond better afterwards. The non-use of probability trees at university in appropriate situations could therefore prove to be one of the causes of the difficulty of students in the secondary-tertiary transition.

Through this example of exercise, I have seen that the recognition of the task by students as part of a succession of types of tasks is complex and not always immediate.

Here are the two main results that I can draw from this analysis. First, during the secondary-tertiary transition, the greater the complexity of linking the task to a type of task, the more difficult it is for students.

Second, the probabilistic interpretation of natural language statements poses difficulties for students, in particular when it comes to identifying the events at stake.

In my future research, I will design and evaluate a teaching aiming to overcome these difficulties.

REFERENCES

Chevallard, Y. (2006). Steps towards a new epistemology in mathematics education. In M. Bosch (Ed.) *Proceedings of the 4th Conference of the European Society for Research in Mathematics Education*, (pp. 22–30), Barcelona, Spain.

Doukhan, C., & Gueudet, G. (in press). Students' difficulties at the secondary-tertiary transition: the case of random variables. *Proceedings of the 11th Congress of the European Society of Research on Mathematics Education*, (pp. 2464-2471), Utrecht, Netherlands.

Durán, P., & Marshall, J. (2019) Mathematics for biological sciences undergraduates: a needs assessment, *International Journal of Mathematical Education in Science and Technology*, 50(6), 807-824.

González-Martin, A., Gueudet, G., Barquero, B., & Romo-Vazquez, A. (to appear). Mathematics and other disciplines, and the role of modelling: advances and challenges. In Durand-Guerrier, V., Hochmut, R., Nardi, E., & Winsløw, C., (Eds.) *Research and Development in University Mathematics Education*, Routledge.

Gueudet, G., & Thomas, M. (2019). Secondary-tertiary transition in mathematics education. *Encyclopedia of mathematics education*. New York, NY: Springer.

Lange, J-M. (2000). Les relations biologie/mathématiques interrogent l'enseignement des sciences de la vie. *Aster 30*,123-142.

Ministère de l'Education Nationale (2011). Bulletin officiel spécial n°8. Mathématiques. Cycle terminal de la série scientifique. Terminale. [Official bulletin #8. Mathematics. Final cycle of the scientific series.] Retrieved from: https://cache.media.education.gouv.fr/file/special_8_men/98/4/mathematiques_S_195984.pdf

Vandebrouck, F. (2008). *La classe de mathématiques : activités des élèves et pratiques des enseignants*. Toulouse: Octares Editions.

Viirman, O., & Nardi, E. (2018). Negotiating different disciplinary discourses: biology students ritualized and exploratory participation in mathematical modeling activities. *Educational Studies in Mathematics*. 10.1007/s10649-018-9861-0.

Vleeschouwer, M.D., & Gueudet, G. (2011). Secondary-tertiary transition and evolutions of didactic contract: The example of duality in linear algebra. *Proceedings of the 7th Congress of the European Society of Research on Mathematics Education*, (pp. 2113-2122), Rzesvow, Poland.

Appendix: English version of the test given to the students

Test

Exercise 1 :
Let be X and Y two individuals whose lifetimes are independent and are such that $\mathbb{P}(X \text{ still living 9 years})= \frac{2}{5}$, $\mathbb{P}(Y \text{ still living 9 years})= \frac{3}{5}$. Calculate the probabilities that :

1. X and Y are still living 9 years;
2. One of the two at least still lives 9 years;
3. X only lives another 9 years;
4. X lives another 9 years knowing that at least one of the 2 will live another 9 years.

Exercise 2 :
You are the director of staff of the Minister of Health. A disease is present in the population, in the proportion of one sick person out of 10,000. The manager of a major pharmaceutical company comes to you to tell you about his new screening test : if a person is sick, the test is positive at 99%; if a person is not sick, the test is negative at 98%. Do you authorize the marketing of this test?

Exercise 3 :
The marketing department of a telephone store conducted a study of its customers' behaviour. He observed that the latter is composed of 42% of women. 35% of women who enter the store make a purchase while this proportion is 55% for men. A person enters the store. We note the events :

- F : "the person is a woman".
- R : "the person leaves without buying anything".

Throughout the exercise, give values approximating the results to the thousandth.

1. Build a weighted tree illustrating the situation.
2. Calculate the probability that the person who entered the store is a woman and leaves without buying anything.
3. Show that $\mathbb{P}(R) = 0.534$.
4. Are the F and R events independent?