FUNDAMENTAL IDEAS IN THE PROBABILISTIC REASONING OF HIGH -SCHOOL STUDENTS IN BINOMIAL DISTRIBUTION ACTIVITIES

Ernesto Sánchez[1], Guadalupe Carrasco[2], and Mariana Herrera[1]
[1] Departamento de Matemática Educativa, Cinvestav-IPN, México
[2] Colegio de Ciencias y Humanidades, Plantel Sur, UNAM-México
esanchez0155@gmail.com

The aim of this research is to identify fundamental probabilistic ideas applied by high-school students in activities related to the construction of the binomial distribution. Activities (including manipulatives and Fathom) were designed with the purpose of enabling, in a specific context of a binomial situation, that the ideas of sample space, combinatory, classical, and frequency approaches of probability, random variable, distribution, and variability are recovered or elaborated by the students. The results show different levels of reasoning with fundamental ideas, difficulties with some of them and, especially, the absence of some inferences. In conclusion, it could be stated that a suitable design of activities for the construction of the binomial distribution encourages students to practice and strengthen their reasoning with fundamental ideas of probability.

INTRODUCTION

This work is focused on the probabilistic reasoning of high-school students regarding four ideas of probability: 1) combinatorial trees and sample space, 2) random variable and binomial distribution, 3) compound experiences, and 4) the relationship between the classical and frequency approaches of probability and variability. All of this is found in a context where the students solve an activity related to the construction of binomial distributions b(x, 3, ½) and b(x, 3, ⅓).

Bakker and Derry (2011) formulate three challenges to statistics (and mathematics) education: 1. Avoiding inert knowledge: knowledge that the students have learned to reproduce but cannot use in an effective way. 2. Avoiding an atomistic approach that shows knowledge as isolated capsules of procedures and concepts. In addition, 3. The challenge of sequencing topics through an approach that promotes coherency from the student's point of view. The way probability is currently taught does not seem to be designed to face such challenges; in any case, we must research what we do and can do within the classroom to face them. In that regard, the proposals that recommend design research are useful. From them, we want to stress the proposal by Cobb and McClain (2004) on the principles of design for the development of statistical reasoning in basic teaching that considers "focusing on central statistical ideas".

As for central ideas, a list of 10 fundamental ideas for stochastic education (Heitele, 1975) was proposed some time ago. More recently, Burrill and Biehler (2011) suggested a list of 7 fundamental ideas for statistics, based on a number of current approaches in the field. For their part, Batanero et al. (2016) have updated Heitele's list of probability ideas. In the present study, we mainly focus on the idea of distribution. This is the third idea in Burrill and Biehler's (2011) list while the stochastic variable is the ninth one in Heitele's work (1975). Batanero et al. (2016) refer to it as a random variable and mathematical expectation.

Considering the educational challenges formulated by Bakker and Derry (2011), fundamental ideas are nodes in a web connected to other fundamental ideas, procedures, and concepts of lower level. One research problem arising is that of exploring and explaining how students manage to construct a conceptual network around a fundamental idea.

We consider that the introduction to the study of the binomial distribution in high school is an opportunity for students to move forward in the review and integration of a number of probability concepts. From an instructional design consistent in the participation of students in the solution to a problem involving basic binomial distributions, which connections and inferences do students make or omit when trying to construct a probability distribution?

THEORETICAL APPROACH

Heitele (1975) considered that fundamental ideas were "ideas which provide the individual on each level of his development with explanatory models which are as efficient as possible and

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

which differ on the various cognitive levels, not in a structural way, but only by their linguistic form and their levels of elaboration." (p. 188). A property of an explanatory model is its capacity to play a role in inferential networks, given that many explanations are nothing more than inferences. In this work, fundamental ideas are not thought to be isolated, but as part of a network of concepts: "Every concept has components and is defined by them." (Deleuze & Guattari, 1991). Besides, fundamental ideas are linked to one another. In mathematics and probability, the most abundant relationships are inferential and these, according to Brandom (a philosopher referenced by Bakker & Derry, 2011), are intertwined with what is conceptual:

> *To grasp or understand (. . .) a concept is to have practical mastery over the inferences it is involved in—to know, in the practical sense of being able to distinguish, what follows from the applicability of a concept, and what it follows from. (Brandom, 2000, p. 48)*

We consider that developing the students' reasoning, whether from statistics or mathematics education, consists in promoting the connection of ideas, particularly the creation of inferences between concepts and notions they know and those that they are acquiring. In the construction of binomial distribution, several relevant ideas on probability converge so that their study involve a range of other, more basic, notions. In addition, more advanced ideas can be developed with binomial distribution.

METHOD

The participants of this research were thirty-four high-school students (ages 17–18) coursing a modality of high school at the National University (UNAM), who had taken a six-month course of probability. In the course, they studied two major themes: descriptive statistics and probability. The subtopics of probability were events and sample space, probability definitions and conditional probability and independence. The research took place while the students took the second semester of the course. They had already reviewed the subtopics concerning discrete random variables, probability distributions, and particularly a little on binomial distribution. Therefore, they were prepared to work on the task proposed in this research work.

The study was divided in four stages: 1) and 4) Applying a questionnaire/task (which doubled as pre-test and post-test); 2) Activities of physical simulation in which questions on worksheets were answered; and 3) Activities of simulation in computer using Fathom software. The questionnaire/activity (see Appendix) formulated two situations ("Random experiment 1" and "Random experiment 2") which could be modeled using binomial distributions b(x,3, ½) and b(x,3, ⅓), respectively. For each of them, the students were asked about aspects of the problem that would lead them to construct the corresponding binomial distribution. The questions dealt with aspects of the problem as parameter of distribution, simple experiment (Bernoulli), tree diagrams and sample space, random variable, and probability distribution. Another question was formulated to find how students used distribution to make a prediction on the frequencies of each value of the variable in 1000 repetitions of the experiments.

In the activities of physical simulation, the students were asked to make 48 repetitions of physical experiments equivalent to the situations of "responses to an exam" stated in the initial questionnaire/activity. After discussing in-group the possible methods to do the simulations, the students used three coins in the situation with p = ½ and opaque bottles (painted in black) with transparent bottleneck and three marbles inside (Brousseau bottle), so that the probability of "Success" were p = ⅓. The students were asked questions before they conducted the simulations for them to identify similarities between the modeled experiment and the original one ("responses to an exam") and predict the number of times each value of the random variable would occur in the different simulations. After completing the simulations, the students were asked questions about the most and least frequent values to determine whether they identified the form of distribution. They were asked to recover the theoretical probability distribution to compare it against the relative frequencies distribution. In the third stage, the students were instructed to carry out simulations using Fathom software and an application developed by the researchers. For each random experiment, two simulations, one of 50 repetitions and another of 1000, were conducted.

Additionally, with the results obtained bar graphs were constructed and analyzed. After a question to focus the students' attention on the similarities between the forms of simulation (physical and on computer), the participants were asked use the worksheets to register the number of times each value of the variable was obtained in both simulations of 50 and 1000 repetitions. They were also asked to see which of the two simulations showed more variation when applying the instruction "randomize", which generates new random samples of the indicated size. Finally, they were asked to work in pairs and write down their observations.

RESULTS

Combinatorial trees and sample space

Combinatorics is a basic tool of probability. Batanero et al. (2016) consider combinatorial enumeration and counting as the third fundamental idea of probability. The concept of combination is used in the construction of the binomial distribution and appears in the algebraic expression of distribution. However, for the construction of simple distributions b(x, 3, ½) and b(x, 3, ⅓) there is no need to know the formula of the combinations, given that there can be a direct count on the description of the sample space. For that reason, the students were asked to construct a tree diagram for each problem.

Making a tree of the situation to describe the sample space demands a trait of combinatorial thought that is often overlooked, which is that of representing potential options and not the real situation. In the first case, 11 students represented the static situation and 17 did so with the second one as follows:

In the representation a and b are the response options. A combinatorial tree is not the mere representation of the situation, as the one before, but a description of a display including all the possible results. This ability of predicting or imagining the possible results, without performing the experiment, is necessary to probability learning. On the other hand, those who manage to construct the tree of two options per question with eight final branches cannot necessarily construct the tree of three options per question with 27 branches. In the study, 14 students found the tree to describe the sample space of the first situation (two options) and only six managed to construct the tree for the second one (three options). For some of them, the obstacle to make the inference was their inability to handle the indistinguishable options of the space (Correct, Incorrect, Incorrect), while others had trouble adjusting the tree with 27 branches in the space provided in the sheet. Finally, some others had trouble with both.

Random variable and distribution

The transition from a probability model to a distribution is done by introducing a random variable. As it is known, this is a function whose domain is the sample space and the counter domain, the set of real numbers. Their nature of function is key to the construction of distribution since the probability of a value of the variable is obtained calculating the probability of the event originating that value; that is, the probability of the inverse image of the value. A distribution of probabilities articulates several concepts of probability and provides a synthetic version of the probability model of a random experience, underlining some characteristic and allowing for its tabular, algebraic, and graphic representation.

Question 4 in the questionnaire defines the random variable: "Consider variable X = 'The number of correct responses'. Describe all the values this variable can take." In Experiment 1, 24 students correctly listed the values of the variable, while 22 students did so in Experiment 2. In question 5, students were asked the probability of the values of the variable; only 15 students in the first experiment and 3 in the second one showed they remembered the relationship between the value of the variable and the event it comes from. The rest of the students proposed values falling in the equiprobability bias and some others provided spurious responses. Several students described the sample space well and, based on it, found the value of the variable; however, they did not register in which event of the sample space each value of the variable was originated.

Compound experiences

There are many ways of obtaining compound experiences from other simpler experiences. A basic procedure to do so is repeating a simple experiment several times. This presupposes the notion of independence in that it is possible to repeat the experiment in similar conditions, so that the probability model of each repetition be the same. In a problem involving the construction of a

binomial distribution, it is suitable to identify the underlying Bernoulli distribution; that is, the event equivalent to "1" or "success" and its probability, in addition to the number or times the experiment is repeated (n) and the event to be calculated. Both probability models, first that of Bernoulli and then the binomial one, are completed with this information.

Question 1 of the questionnaire "What is the probability that a question is answered correctly?" was found ambiguous. The researchers' aim was for the students to identify the Bernoulli experience, namely the experience of responding a question whose probability is simply ½ since a question only has two options of response. Nevertheless, 10 students in situation 1, and 7 in situation 2, interpreted they were asked the probability of correctly responding a question of three because they had been previously shown a situation of a multiple-choice exam with three questions each. Besides, they made the mistake of interpreting the situation as "a success in three tries" and responded ⅓ instead of ⅜. This evidences the importance of the difficulty when observing the binomial structure of the situation and identifying the underlying Bernoulli experiment.

Frequentist approach of probability and variability

The repetition of independent and identically distributed experiments is a resource to provide meaning to the concepts of theory of probability and a bridge towards its applications. In addition, within this resource, the relative frequency is the key concept together with its property of stabilizing around a number. The technological resources and the idea of simulation allow recreating these ideas; and such behavior can be observed with distributions and not only isolated events. Simpler binomial distributions are ideal for this and since they have a few values, punctual and global behaviors are observed. A critical element that can be integrated thanks to simulation is considering the variability of the relative frequencies and the corresponding behavior of the expected (absolute) frequencies.

As we said before, 15 students managed to obtain the probability distribution b(x, 3, ½) while only three obtained b(x, 3, ⅓). From them, only three cases provided signs that the students remembered theoretical distribution when describing the empirical distribution obtained through simulation, but without making the relationship between them explicit. In the context of the simulation, when asked to calculate probabilities, the rest of the students chose to assign frequentist probabilities without referencing the theoretical ones. A consequence of the simulation activities was that 12 and 17 students in the first and second experiment, respectively, proposed probability distributions defined with relative frequencies.

Regarding the prediction problem, we expected the experience to trigger a sense of variability in the students that would be evident in proposals identifying the expected frequencies, but without modulations as "approximately", "around" or with proposals as "237, 512, 251" that implicitly reflect both the knowledge of expected experiences and variability. In this regard, the simulation activities had little influence on the students. There were no significant changes in the way they responded this question before and after the activities.

CONCLUSION

The binomial distribution is ideal to help students create relationships and make inferences between different fundamental ideas of probability. To do so, we should characterize some key elements that represent obstacles students must overcome when constructing such distribution. In the following, we will deal with some of those aspects, stressing the fact that some of the students faced difficulties. Nevertheless, we should also consider that, in all the cases, there were students who managed to successfully do what they were asked to. An important aspect in the binomial construction is referred to combinatorics. From our observations, the binomial is related to the combinatory. In addition, it must be established that combinatory procedures, in binomial situations, are meant to describe potential situations, what is possible and feasible and not real situations. Besides, the problem of generalization in the procedures faces the difficulty of handling indistinguishable objects and the exponential growth of arrangements and combinations as the objects in play increase. This demands giving up the concrete aspects of such procedures (trees) to operate more conceptual ones (rule of product, commutations, combinations, etc.).

The name random variable hides its nature of function and the description of the values it takes seems to be enough, yet it is not. On the contrary, we should perceive that each value of the

variable has its origin in an event. The observations found in this work indicate that many students see unidirectional relationships that go from elements of the sample space to values of the variable. Still, they do not construct the events of the sample space, which give rise to such values. In other words, they do not see event-value relationships of the variable. In consequence, the partition of the sample space, which would allow them to calculate the distribution probabilities, is absent. This shortcoming leads to a rupture, so that there were students who adequately described the sample space, yet they provide a spurious distribution based on the equiprobability bias. Then, we must relocate the idea of random variable in teaching in order to better understand its nature of function and its key connection between the concrete and abstract mathematical models, which are distributions.

Laplace's definition of the probability of an event as the quotient of possible cases divided by favorable cases might become a scheme that undermines the perception of binomial situations. In the case of the problem of responses to an exam that we analyzed, some students organized the situation with a sample space size 3, corresponding to each question. In these cases, they did not start modeling by identifying the Bernoulli situation. This indicates the importance of working harder on creating random experiences and, particularly, considering their repetition.

Making students establish the relationship between classical probability and frequentist approach of probability is a challenge. In our experience, most of the students know both approaches but see them in an isolated manner. In the context of simulations, they have no trouble proposing a distribution based on relative frequencies provided by a simulation using software; however, most of them do not associate it with the distribution found in the previous tasks. When students are located in the context of the classical definition, they consider that what they see in the simulation has no consequence. Then, in prediction problems, they anticipate the expected frequencies (250, 500, 250), without modifying their prediction after observing the variable in the simulations.

In the above, we sought to stress aspects that should be considered in the design of instructions to learn binomial distributions and help to prevent the knowledge students acquire during the learning process of the subject from accumulating in inert knowledge. The observations we have made indicate elements on which we should pay attention so that students can construct a coherent system from their point of view with the fundamental ideas of probability.

REFERENCES

Bakker, A. & Derry, J. (2011). Lessons from Inferentialism for Statistics Education. *Mathematical Thinking and Learning, 13*(1 & 2), 5-26.

Batanero, C., Chernoff, E., Engel, J., Lee, H., & Sánchez, E. (2016). *Research on Teaching and Learning Probability*. Springer International Publishing.

Brandom, R. B. (2000). *Articulating reasons: An introduction to inferentialism*. Cambridge, MA: Harvard University Press.

Burril, G. & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In C. Batanero, G. Burrill, C. Reading (Eds.), *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education* (pp. 57-69). New York: Springer.

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking (pp. 375–396).* Dordrecht, The Netherlands: Kluwer Academic Publishers

Deleuze, G., & Guattari, F. (1991). *What is philosophy?* New York: Columbia University Press.

Heitele, D. (1975). An epistemological view on fundamental stochastic ideas. *Educational Studies in Mathematics, 6*, 187-–205.

APPENDIX

Questionnaire (Pre and Post-test)

Random experiment 1. A multiple-choice exam consists of three questions, each of them with two options of response, one of them is correct. The student responds each question by randomly choosing one of the options. The options chosen by the student are observed.

1. What is the probability of responding a question correctly? Explain your answer.
2. Describe all the possible, different forms of responding the exam (use a tree diagram).
3. How many different results does the Sample Space of the experiment have?
4. Consider the variable X = "The number of correct responses." Describe all the values that this variable can take.
5. Do the following:
   a) What is the probability that the variable takes the value 0?
   b) What is the probability that it takes the value 1?
   c) What is the probability that it takes the value 2?
   d) What is the probability that it takes the value 3?
6. Based on the previous responses, complete the following table [Discuss which table is the most appropriate]:

Values of X					Sum
Probability					

7. 7. If the student passes the exam when responding at least two questions correctly, what is the probability that the student passes the exam?
8. 8. If 1000 students responded the exam and all of them responded randomly
   a) How many of them would guess zero questions correctly? Explain your answer.
   b) How many of them would guess only one question correctly? Explain your answer.
   c) How many of them would guess two questions correctly? Explain your answer.
   d) How many of them would guess three questions correctly? Explain your answer.

Using the responses, complete the following table:

Values of x	Frequencies
Sum	

Random experiment 2. A multiple-choice exam consists of three questions. Each question has three *options*, one of which is correct. A student responds each question by randomly choosing one of the options. Respond the following questions: [Same questions as in Random Experiment 1].