AN EXPLORATORY STUDY OF STUDENTS' DIFFICULTIES WITH RANDOM VARIABLES

Blanca R. Ruiz Hernández, José Armando Albert Huerta
ITESM, Campus Monterrey, México
Carmen Batanero
Universidad de Granada, España
bruiz@itesm.mx

In this research we approach a fundamental stochastic idea. The random variable is based on other mathematics and probabilistic concepts and, in turn, is the support of many probability and statistics subjects. In this paper, we present some results from an exploratory study carried out with two university students. The aim was observing the difficulties the students face when they try to solve a problem that involves the concept of random variable.

INTRODUCTION

In his study about fundamental stochastic ideas, Heitele (1975) included the random variable in his list of ten fundamental ideas in statistics and probability courses. However few researchers have explicitly focused on an analysis of students' difficulties in learning this concept that is basic in the development of probability distributions, central limit theorem, hypothesis testing and other topics common in learning of probability and statistics.

Random variables, as well as probability and distribution functions, are powerful tools that introduce mathematical analysis in probability (Batanero, 2001). However, the concept is neither easy to learn or to teach. The definition is simple only in appearance: "a rule that assigns exactly a numerical value to each outcome in a single sample space," because the concept depends on an understanding of random phenomena, random events, event operations and probability. It is also required that the inverse of an interval in the random variable transformation be a measurable set for the rule to be a random variable. Only on this case, can we speak about a probability distribution that determines the random variable in a one to one function and vice versa. This means that random variables are not just based on probabilistic tools, but also on mathematical tools usually studied as part of deterministic mathematics in Secondary School. Finally, random variables are linked with the mathematical modeling process, which serves to relate random real phenomena to the mathematical context of probability distributions (Ruiz, 2004).

Moreover, the concept of random variables is not related to students' intuitive experiences in formal education, where it is often presented only as a preamble to probability distributions (Oseguera, 1994). Miller (1998) and Ortiz (2002) suggest the need to revise some textbooks where the study of random variables is totally or partially omitted, or the introduction of the concept involves errors that might influence students' understanding. Tauber (2001) deduced an important lack of connection between the study of probability models and empirical data in textbooks. Thus, the relevance of doing research on learning difficulties in random variables is justified on both epistemic and didactic reasons, which are inseparable.

METHODOLOGY

We carried out a *cognitive exploration* to identify the constructive processes for the idea of the random variable with a pair of University students when they were solving a problem in which the concept was involved. The students had completed their first mathematics course (calculus) and they had not taken any previous statistics course at University. They were selected from among the best students in the course (according to their scores in previous courses) and because they were communicative. Other students helped in piloting trials of the activity. All of them volunteered to take part in the experience.

Cognitive exploration is a methodological tool widely used in clinical research (e.g., Schoenfeld, 1985; Nemirovsky, 1993). It is a clinical qualitative technique that serves to explore how a small group of two or three students solves a proposed problem without the help of their teacher, but in his/her presence. The teacher's role is reduced to clarifying or deepening the ideas

expressed by the students either to him/herself or to the research. The students solve the proposed activity by consensus, expressing their ideas aloud and arguing their positions. The teacher's questions to the students are conditioned by the students' responses to the problem, as well as by a previously prepared questionnaire.

Our goal in the exploration was to describe some of the difficulties that students face when dealing with random variables and how random variables relate to other concepts. We were also interested in exploring their conceptions and the ideas they develop in solving a problem situation that involves random variables. The design of the activity was supported by previous hypotheses about how students would face the problem and their possible difficulties, and on some didactic variables we considered useful to explore.

Hypotheses

The following hypotheses were used to guide the researcher during the interviews:

1) *Lack of perception of randomness in the problem*. Students' perceptual difficulties as regards randomness have been described by many authors, for example, Falk and Konold (1997) and Batanero and Serrano (1999).
2) *Tendency to "algebrize" and not contextualize the procedures related to random variables*. One main reported bias related to random variables is the predominance of deterministic thinking over probabilistic reasoning (Nardecchia and Havia, 2003). We anticipated that the mathematical modeling process would be reduced to working with the mathematical model, and that the model generation and the interpretation of results would be forgotten.
3) *Surprise when working with functions in a probabilistic context*. Our conjecture was that the students would be surprised to work with a dependent variable in a function that is so intangible as the probability distribution, which is not a physical magnitude, but an abstraction.
4) *Difficulties with the formal notion of random variable*. Heitele (1975) suggested that the intuition of random magnitude appears earlier than the intuition of random experiment. We therefore expected students to have an appropriate intuition of random variables, even when having some problems when working at a formal level.

Design of the Activity

The students were asked to solve Problem 1 (shown below) that was given to them in a questionnaire with some additional open questions. The questions guided the solving process and led the interviewer to explore the students' reasoning. The total time spent in solving the problem and successive interviews was three and a half hours, which were videotaped for later transcription and analysis.

Problem 1: There is a celebration of Children's Day in a town. The public relations department in a factory is organizing a lottery for the workers' children. Each worker has equal likelihood to win. The winner will be given theatre tickets for him/her and his/her whole family. However, the tickets must be bought some days in advance of the celebration.
The company manager has to decide how many tickets to buy so that there are enough tickets for the whole family of the winner but not an excessive number of extra tickets. The distribution of the number of children per worker is given in the following table.

Number of children	0	1	2	3	4	5	6	7	8	9
Number of workers with these children	16	22	33	45	31	20	12	9	7	5

The students solved the problem in the presence of their mathematics teacher and a researcher. The questionnaire was centered on the following questions:

- *Part I. Introducing the problem*. In this part we introduced the problem. We questioned the students about the classic probability notion (given by the Laplace rule as the quotient between the number of favorable and possible cases) and asked them to define the variables that enter into the probability function.

- *Part II. Main questions in the problem*. The students were requested to make a decision and explain their position. They should also define their ideas of risk and cumulative probability. The sample space and random event concepts were also explored.
- *Part III. Change from tabular to graphical representation.* There was also an exploration of the random variable meaning and interpretation. At the end, we asked the students to compare stochastic and deterministic situations.

ANALYSIS OF THE RESULTS

We present a summary of our results, organized around the three main variables in our study.

Randomness

The students accepted the randomness linked to the problem. At the beginning of the activity the students did not identify a probability problem as a mathematical problem. They even suggested the problem was just ethical. Their initial reticence to accept uncertainty was substituted by assuming the randomness in the problem. At the same time, they related randomness to probability and interpreted probability as a measure of the uncertainty in the situation, which permitted them to get a solution through various notions, such as probability distribution, mode and median. Possibly, this linking of probability and randomness led students to accept the randomness in the solution of the problem and therefore to provide a solution. Since they were able to use probability to decide how many tickets to buy, they found a way to be ethically equitable, in spite of uncertainty. Their acceptance of uncertainty was linked with the certainty that their solution was, in some way, equitable. Given that certainty was not possible, they sought some form of fairness. It is important to point out that the notions of fairness and equitable situations are the first steps in building a probabilistic intuition (Cañizares *et al*., 2003).

Later, in the third part of the questionnaire, the students related the randomness of the situation to probability instead of relating it to the random variable. The probability value would be random for them since probability was computed from data and not from a formula and they linked randomness to their inability to compute probabilities for the random variable. Thus, the students considered probability to be 'unpredictable' because it depended on the frequency of workers with a given number of children. That is, since two different factories might have different workers and different number of children in the families they considered it impossible to obtain the same data distribution in another factory. They suggested these were *'"human," not "mathematical" data.'*

The Probability Notion

The students had an appropriate understanding and correctly computed classic probability (defined by the Laplace rule). However, they did not find it so evident that probability was the dependent variable in the probability function. The number of workers was more clearly viewed as a dependent variable by them. We will report on various aspects of this view in the following.

First, the students found it is difficult to move from the context of the problem to the mathematical context. The variable 'number of workers' was closer to the context of the problem for them than the formal probability, described through the probability function, which is an abstract mathematical object. Their interpretation of classical probability possibly caused them "to see" probability, as two numbers instead of just one (probability was for them '30 out of 200,' instead of $\frac{30}{200}$). Consequently they preferred to work with the number of workers (30) instead of working with probability (0.15 or $\frac{3}{20}$), in spite of their correct interpretation of probability both as a fraction and as a decimal number. That is to say, the knowledge that was successful in the classic probability context was an obstacle to construct the idea of formal probability. Similar didactical obstacles have been described in the learning of rational numbers.

The fact that two relationships and four variables (one of which is, in fact, a parameter) intervene in obtaining the probability function hindered the students' acceptance of formal probability as the dependent variable in the probability function. Little by little, they discovered

and differentiated the two relationships: a) computing probabilities through the Laplace rule and b) the frequency histogram of the number of children per family. They, therefore, also differentiated the four variables involved in the situation: number of workers in each category (frequency of an event in the sample space), total number of workers (size of the sample space), each worker's number of children (value of the random variable) and probability of having a number of children. When they realized that the number of workers was needed to compute classical probability, and at the same time it appeared in the frequency histogram, but playing different roles in both relationships, they realized that the number of workers was different from probability.

The students did not accept that mathematics serves to represent reality and they detached the problem context from the mathematical representations. This separation was emphasized by the discrete context, because, as they argued: '*mathematics has always been continuous.*'

The Random Variable Notion

The use of the random variable as a correspondence rule was natural for the students. They even suggested that the "number of children per worker" related the random phenomenon to the solution by using the cumulative probability and by interpreting the probability distribution. That is, even when the result of the lottery was the winner's name, and we could be interested in any data about the worker (age, etc), we in fact were interested in the number of children. The "worker's number of children" was the variable related to probability in the probability distribution. Presenting the random variable as a correspondence rule (that is to say, the "rule" that links the events with a real number) is a rich possibility to be explored.

Again, we found problems in relating the mathematical and problem contexts. The students realized that the random variable in the probability function graph represented the worker's number of children and that, in case the experiment was repeated a large number of times, the results would not appear in an established order. For this reason, they believed that the uncertainty of order in the results could be translated to the abscissas-axis. They then, suggested that we might represent numbers on this axis in any order without affecting the analysis. Again this difficulty was linked to the lack of distinction between model and reality (Henry, 1997).

The real numbers, which are assigned to each event by a random variable, play two roles in this process: they are a part of the mathematical model in the first stage of modeling (when the random variable is defined) and a part of "reality" at another stage (when the probability distribution is defined). In the first stage, "reality" is given by the random event to which we assign a number and the "mathematical model" is the random variable assigning the numerical value to each outcome of the sample space. In the probability distribution, "reality" is a real number, inverse of a random variable (independent variable) and probability is the "mathematical model." The students did not accept the random variable as a mathematical model when they failed to assign real number properties to the number of children that for them continued being part of reality. The numbers of children were seen as labels of "reality" and not as real numbers.

Thereafter, they linked the uncertainty in the probability function to 'probability' rather than to the 'number of children' because they wanted to generalize the probability function. They were uncomfortable working with a function that was useful only in this situation and would change in other problems. They did not realize that, in this context, the probability was fixed for each number of children and what was uncertain was the number of children the winner. Since we did not have an equation and the data would change depending on the context of each factory, the probability would change. However the number of children would continue taking the same values 0, 1, 2, 3... So they concluded that the number of children was not 'unpredictable.' We think this problem was caused by the students' previous mathematics courses, where graphics are taught without a context and led students to think a function can serve in various contexts and it does not change in different contexts.

These problems were also related to the students' difficulty in expressing themselves in mathematical language. Thus, for them a graph did not have to be related to a problem context, it was something mathematical living by itself and did not have practical or interpretative goals. Similarly, the students interpreted the idea of probability as a quotient, because it was easier.

They also tended to speak of "number of workers" because this expression was closer to the context than probability.

CONCLUSIONS

Lack of Perception of Randomness in the Problem

We observed the students' difficulty in perceiving randomness in the first stages of their work, where they insisted in giving a deterministic solution to the problem, although later this difficulty was overcome. Several times the students did not link mathematics to the problem context, showing a conception of mathematics as something that is justified by itself, without having a relationship to a situation in real life. This is the first ontological obstacle in the process of modeling. Then, they did not link the algebraic work, which is used in other mathematical topics, to probability because they considered it is too real to be analyzed by mathematics. This separation was shown when they found it strange not to be able to assign an algebraic expression to the function. That is to say, they expressed an algebraic conception of dependence (Estepa, 1994), which is consistent with believing that all relationships of dependence between variables should be given by a formula. Moreover, they suggested that the probability function was not controllable because the value of the dependent variable came at random. This might be an epistemological as well as a didactical obstacle since the rest of mathematics is only studied from a determinist point of view.

Tendency to "Algebrize" and not Contextualize the Procedures Related to Random Variables

For these students, mathematical deterministic functions were different from this probability function because the value of the dependent variable (probability) was not predictable as a function of the independent variable (number of children). They thought functions were general and not tied to a particular situation and believed that obtaining an equation for the probability distribution made no sense. Though it was true that the distribution function for this particular problem was not applicable to other problems, this did not imply that solving the problem was irrelevant.

Surprise when Working with Functions in a Probabilistic Context

The students were surprised in associating the function concept to the probability idea, though they found natural the use of frequency tables and probabilities. They also had difficulty in working with a discontinuous function (step function), where the independent variable was discrete, and where the domain was so restricted. They related the graph of a function to the continuity of the drawing, possibly because this was the only type of function with which they had previously worked. Thus, they did not find contradiction in having a decimal value for the independent variable (the number of children). That made no sense in the problem context although it was in agreement with the didactic contract established in reference to functional graphics. They believed that the graph of a function would be the same independently of having a discrete or continuous independent variable. They gave a similar interpretation to the domain even when it was not possible to compute the probability for a fraction of a child.

Another difficulty was related to the modeling of random phenomenon. The variable 'number of children' played two roles in the relation between mathematical model and reality at the different levels of the mathematical model. In one of them, the 'number of children' was a real number with mathematical properties; while, in the other, it served to link a random phenomenon ("reality") with mathematics. This was not perceived by the students who did not consider the need to order values of the *x*-axis in the graphical representation of the probability distribution because in repeating the random experiment many times, the random variable would take its values in varying order.

Difficulties with the Formal Notion of Random Variable

At the beginning of their work, the students confused probability with the number of workers, but later they intuitively visualized the probability distribution as a compound function. In that compound function the number of children plays a double role in each of the functions that compose it: dependent variable in the inverse application of the random variable; and independent

variable in the probability function. Eventually they were able to pass from one function to the other.

We note the epistemological complexity of random variables, reflected in the difficulties that the students had when solving the problem. They linked the idea of a random variable to complex probabilistic concepts, related it to a mathematical modeling process and connected it to other stochastic and non stochastic mathematical tools.

ACKNOWLEDGEMENT

Research supported by the grants FQM126, Junta de Andalucía and SEJ2004-00789, MEC; Madrid.

REFERENCES

Batanero, C. (2001). *Didáctica de la Estadística*. Granada: Grupo de investigación en educación estadística. España: Universidad de Granada.

Batanero, C. and Serrano, L. (1999) The meaning of randomness for secondary students. *Journal for Research in Mathematics Education*, 30, 558-567.

Cañizares, M. J., Batanero, C., Serrano, L., and Ortiz, J. J. (2003) Children's understanding of fair games in *CERME* 3.

Estepa, A. (1994) *Concepciones Iniciales Sobre la asociación Estadística y su Evolución como Consecuencia de una Enseñanza Basada en el Uso de Ordenadores*. Tesis Doctoral. Universidad de Granada.

Falk, R. and Konold, C. (1997). Making sense of randomness: Implicit encoding as a basic for judgment. *Psychological Review*, 104, 310-318.

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

Henry, M. (1997). Notion de modéle et modélisation dans l'enseignement. In *Enseigner les probabilités au lycée*, (pp. 77-84). Reims: Commission Inter-IREM.

Miller, T. K. (1998) The random variable concept in introductory statistics. In L. Pereria-Mendoza, L. S. Kea, T. W. Kee, and W-K. Wong (Eds.), *Statistical Education - Expanding the Network: Proceedings of the Fifth International Conference on Teaching Statistics*, Singapore, (1221-1222). Voorburg: The Netherlands: International Statistical Institute.

Nardecchia, G. and Hevia, H. (2003) Dificultades en la enseñanza del concepto de variable aleatoria. Trabajo presentado en el *V Simposio de Educación Matemática*. Chivilcoy, Argentina.

Nemirovsky, R. (1993). *Symbolizing Motion, Flow, and Contours: The Experience of Continuous Change*. Thesis (Ed. D.) USA: Harvard University. Graduate School of Education.

Ortiz, J. (2002) *La Probabilidad en los Libros de Texto*. Tesis doctoral. España: Grupo de Educación Estadística de la Universidad de Granada.

Oseguera, F. (1994) *El Concepto de Variable Aleatoria en el Contexto del Currí*culo. Análisis y Alternativas. Tesis para obtener el grado de maestría. México: CINVESTAV-IPN.

Ruiz, B. (2004). *Exploración Cognitiva Sobre la Variable Aleatoria en una Situación de Modelación*. Trabajo de investigación tutelado. España: Universidad de Granada.

Schoenfeld, A. (1985) *Mathematical Problem Solving*. New York: Academic Press.

Tauber, L. (2001). *La Construcción del Significado de la Distribución Normal a Partir de Actividades de Análisis de Datos*. Tesis Doctoral. España: Universidad de Sevilla.