Probability Calculus and Connections between Empirical and Theoretical Distributions through Computer Simulation

Santiago Inzunsa
Facultad de Informática, Universidad Autónoma de Sinaloa
Culiacán Sinaloa México
sinzunza@uas.uasnet.mx

Summary

In this article, we analyze results of a study with 15 university students about the way binomial and normal distributions are constructed and how probabilities are calculated through computer simulation. The connection with theoretical probability distributions, the analysis of low probability values and the effect of the number of simulations on the precision of the results are highlighted. We chose the software Fathom for its cognitive potential and flexibility to simulate random events. The results indicate that the simulation of distributions was relatively simple for the students, who showed an understanding of the effect of the number of simulations on the precision of the results, an ability to interpret the calculated probabilities with more facility than when used probability tables and formulas, and seemed to overcome their misconception that values that do not appear in probability tables are impossible to happen.

KEYWORDS: Probability distributions, Empirical Probability Distributions, Computer Simulation.

Introduction

When we study random phenomena, we are frequently interested in certain numerical values that are associated to or that are determined by some of its results. These values constitute a random variable because the phenomenon is determined by chance. All the possible values of a random variable and their respective probabilities constitute a *probability distribution*. For example, we consider a random experiment that consists of throwing two coins and observing the results that appear. If we denote the random variable that represents the number of tails by X, its probability distribution can be expressed by means of the following expression:

$$\mathrm{P}(\mathrm{X}) = \begin{cases} \frac{1}{4} & x = 0 \\ \frac{1}{2} & x = 1 \\ \frac{1}{4} & x = 2 \\ 0 & \text{elsewhere} \end{cases}$$

Therefore, the probability distribution of a random variable contains information about the possible values that the variable can take as well as their probabilities, which is very important when we study the behavior of random phenomena. In addition, many random phenomena show similar characteristics and generate random variables with the same type of probability distribution. Therefore, knowledge of the probability distributions of random variables of common phenomena eliminates the need to solve the same probability problems every time they appear.

In addition, the probability distributions are important because:

- An understanding of probability distributions is crucial for the comprehension of the statistical models ubiquitous in scientific research.
- The concept is equally central to participation in the public forum as an informed citizen.
- Without the concept of distribution, learners cannot truly understand how events can be both unpredictable and constrained.
- Probability distributions stand at the interface between the traditional study of probability and the traditional study of statistics and, thus, afford an opportunity to make strong connections between the two disciplines (Wilensky, 1997, p. 175).

Probability distributions are theoretical models for modeling a diversity of real life situations where chance takes place. They also allow a description of the variation due to sampling - particularly the normal distribution - with which important concepts of statistical inference as the central limit theorem, confidence intervals, and hypothesis tests are linked.

This way, probability distributions are, and likely will always be a major part of a first statistics course, because probability distributions, like the normal curve are part of the vocabulary for communicating basic ideas in different sciences. While they might not be so frequently used as the statistical concepts of averages and frequencies, they clearly extend beyond the classroom and consequently deserve a special place in the statistics curriculum (Cohen & Chechile, 1997, p. 254).

Nevertheless, in our teaching experience, as well as in communicating diverse results of research (e.g. Tauber & Sanchez, 2002) we have observed that probability distributions constitute a complex and abstract concept for many students. Different causes have an influence in it; amongst all we mention the following:

A stochastic experiment may be analyzed differently:

- A narrow local perspective is to judge the probability of a single event; e.g. E (a six in rolling a die). The event is concrete and named; it occurs or it fails to occur in one trial of the experiment. One may count its absolute or relative frequency. Problems in understanding the probability of this event start when one asks what a probability information of 1/6 means. Is it right to have no six in a single trial, or in three successive trials? A probability always conveys hypothetical not-easy-to-understand information about a real situation.

- A broader perspective is to describe the distribution of all outcomes of an experiment. This is a global perspective to the same phenomenon; e. g. in rolling the die one may be interested in all outcomes of the die and model it by a uniform distribution on the numbers $\{1, 2, 3, 4, 5, 6\}$. One may also investigate the distribution of waiting times for the first “six” – the geometric distribution on the set $\{1, 2\ldots, n\}$. Problems in understanding multiply compared to the local perspective of one event: Many events may be thought of (“all” subsets could be considered); the set of outcomes may be infinite, and even be an interval or the whole set of real numbers. And the calculation of probabilities has to be done by (possibly infinite) sums or by integrals. Do not forget about the difficulties in interpreting only probability of one event – here we have many, not named or explicitly addressed events.

Furthermore, these probability distributions as mathematical entities are described by parameters, which influence the shape of these distributions. What does a specific shape of a distribution mean? And which implications does a specific parameter have upon shape or upon the mean of the distribution? Furthermore, the calculation of probabilities in the form of intervals (a, b) is a technical task which surmounts to calculating a definite integral. The possibility to avoid using such concepts from calculus by basing the calculations on tables increases the calculation difficulties even and does not alleviate the approach. It involves technicalities such as standardizing a and b (with the normal distribution), using symmetries to get the tail probabilities, etc.

The wide variety of distributions that are used to model different phenomena of reality contributes to further confusion (parameters of those distributions get a different meaning). Some of these distributions, even discrete distributions may be approximated by the normal distribution – why (the theorems behind involve sequences of distributions and limits thereof – usually even in the formulation of the relations beyond the scope of statistics courses, not only for applied introductory courses) and how is it performed? One may see that mathematical concepts involved and technicalities to calculate the required probabilities are highly complex. No wonder students stick to inflexible step 1, step 2, etc. strategies to solve the tasks without trying to understand what is going on. Exam papers for decades have shown their relative inability to master the technical part to derive the solutions.

Another aspect is what a distribution implies for a phenomenon, which is modeled by it. Each of the distributions may be linked to a characteristic modeling idea behind. The normal is an approximation to any distribution, if the variable is or might be thought of as a sum of other variables (e. g. the chi-squares are always easy to approximate by the normal if the degrees of freedom are higher – which corresponds to more parts in the sum). Needless to say that such modeling ideas which are referred to here, are not easy to explain to students without the development of the complex mathematical background.

One more aspect is what a distribution could mean, what it could convey for real data, which should be modeled by that distribution. Remember the difficulty to understand the relationship between the probability of an event E and its relative frequencies in repeated trials. For a distribution this cannot be solved mathematically. It cannot be done by a reference to students' experience either – no one "sees" a distribution acting to generate a specific sample repeatedly and may thus learn how strongly the results vary "around" the distribution. Such guided experience, however, is possible today by the capacity of computers and the simulation techniques. However, this is not generally accepted by statisticians teaching their university students.

This is, the deductive approach based on random variables defined in a probability space requiring certain mathematical knowledge that many university students do not have, since the random variables constitute a concept neither easy to learn nor easy 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 understanding random phenomena, random events, event operations and probability (Ruiz et al., 2006).

Many students, particularly more advanced students, learn the symbols, realize the operations with the involved events and by means of axioms and properties of the probability, they calculate probabilities of the random variables with tables or formulas, but often without an understanding of the concepts involved and their limitations.

Since the emergence of diverse computational statistical packages, a different approach based on the simulation principle to the study of probability distributions is possible (Inzunsa & Quintero, 2007). In the same direction, Wilensky (1997, p. 175) considers that in a typical course in probability and statistics, students are exposed to a standard library of distributions and associated formulae, but do not have a chance to construct these distributions and understand what "problems they are trying to solve". The availability of new computational packages of software offers learners the possibility to construct these distributions as patterns emergent from probabilistic rules. Through these connections, learners can make connections between probabilistic descriptions of discrete phenomena and statistical descriptions of the ensemble.

The use of this connection between empirical and theoretical results for teaching purposes is advocated by various statistics educators and researchers. For instance, Fischbein and Gazit (1984), pointed out the importance of teaching activities where the calculus of theoretical probabilities and observed frequencies are related, as a way to enhance the development of efficient probabilistic intuitions. Pfannkuch (2005) proposes a pedagogical frame to introduce connections between probability and statistics in a gradual way for the students to see how data can be modeled by probability distributions. On the other hand, Batanero, Henry and Parzysz (2005) affirm that by incorporating computer technology in statistical education, probability can be seen as a theoretical tool that may be used to study problems that emerge from statistical situations.

In recent curricular reforms, the teaching of statistics and probability is addressed in this orientation. For example, in the Principles and Standards for School Mathematics (NCTM, 2000), the Standard of Data Analysis and Probability for levels 9-12, recommends emphasizing understanding of probability distributions and using simulations to construct empirical distributions, "In the higher levels, students may apply probability concepts to predict the probability of an event and building up probability distributions for simple sample spaces" (p. 336).

The main hypothesis of this study is that the students can develop a better understanding of the concepts involved in probability distributions – in particular for the binomial and normal distribution – in a teaching environment based on computer simulation rather than by the traditional approach based on the use of tables and formulae. We were especially interested in investigating aspects such as probability calculus, connection between empirical and theoretical results and the influence of the parameters in the shape of the distributions.

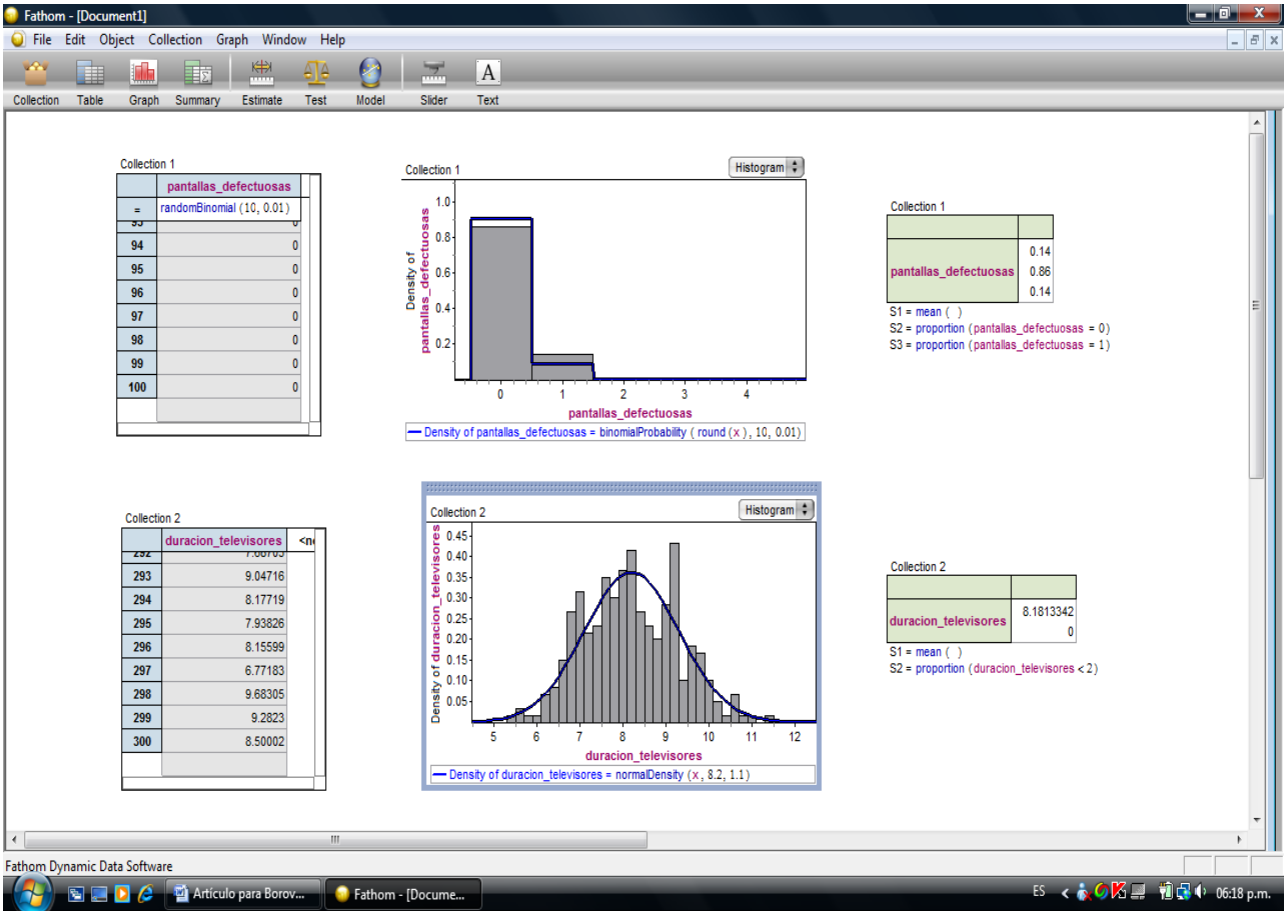
This way, the study compares two of the methods of teaching probability distributions, traditionally by means of formulae and tables and by analyzing results from sampling by computers. The author will focus on students' experiences, how they perform, how they solve tasks with the help of simulations, and how this influences their understanding of the concepts of probability.

Fathom and Probability Distributions

In this research work, we have selected Fathom (Finzer, et al., 2002) instead of other statistical software packages or spreadsheets like Excel, due to its flexibility to perform simulations, its dynamic representations allowing to visualize any changes at graphical and numerical levels when some parameters of the distribution are modified, and because of the facility to superpose the frequency distribution generated by the results of a simulation and the theoretical probability distribution.

To simulate the probability distribution of a random variable in Fathom, first an empty table of cases is selected, the name of the variable (attribute) is defined, and the appropriate command (random number generator) is selected from the formula editor; then the number of cases or samples to simulate is specified. The software places the results obtained in the case table in the order that they were generated and later the results can be transformed into a graphic (histogram) to visualize their distribution.

The probabilities are calculated by means of a formula that is introduced in the summary table. As the number of simulations increases, the histogram of the frequency distribution tends to look like the theoretical probability distribution. Thus, the theoretical results of probability distributions can be compared to the empirical data generated through simulation. In addition, this approach allows exploring concepts that underlie probability distributions, such as the effects of parameters on the shape of a distribution and the effect of the number of simulations on the precision of probabilities. Furthermore, it provides other representations to the unique numerical representations (graphical and symbolical) that are used in probability tables. In case of the binomial distribution, the values of the variable (number of "successes") in each sample of size *n* taken from a population with parameter *p*, are generated by means of the command *randombinomial (n, p),* whereas in case of the normal distribution, the values are generated by means of the command *randomnormal (m, s)* (see Figure 1).



Collection 1

	pantallas_defectuosas
=	randomBinomial (10, 0.01)
94	0
95	0
96	0
97	0
98	0
99	0
100	0

Collection 1

Density of pantallas_defectuosas = binomialProbability (round (x), 10, 0.01)

Collection 1

pantallas_defectuosas	0.14 0.86 0.14

S1 = mean ( )
S2 = proportion (pantallas_defectuosas = 0)
S3 = proportion (pantallas_defectuosas = 1)

Collection 2

	duracion_televisores	<n
293	9.04716	
294	8.17719	
295	7.93826	
296	8.15599	
297	6.77183	
298	9.68305	
299	9.2823	
300	8.50002	

Collection 2

Density of duracion_televisores = normalDensity (x, 8.2, 1.1)

Collection 2

duracion_televisores	8.1813342 0

S1 = mean ( )
S2 = proportion (duracion_televisores < 2)

Figure 1. Simulation of data for a binomial and a normal distribution with Fathom.

The principal actions of Fathom to simulate a probability distribution (table cases, graphics, summary tables and the formula editor) are easily accessible to the user.

The Computer as Cognitive Tool

Reviews of the literature about computer-based learning reveal that computers may support cognition in different ways. However, Pea (1987) observes that there is a dominant tendency in mathematics education in which computers are considered simply as amplifiers of mind (amplifier metaphor). From this perspective, computers are considered as tools that allow carrying out tasks more easily and faster than in a paper and pencil environment, but without a qualitative change in the way of doing it. Another perspective considers computers as cognitive tools. From this perspective, computers -when used appropriately- not only allow to amplify the mind, but also have the potential to generate structural changes in the mind of the users (reorganizer metaphor) through a reorganization and transformation of the activities that they carry out.

Pea (1987, p. 91) defines a cognitive tool as "any media that helps transcend mind limitations, thinking, learning and problem solving activities", while Jonassen (1994), considers a cognitive tool as any mental or computational dispositive that helps, guides and extends the users' thinking processes. Particularly, in the case of computers, they constitute an extraordinary and powerful cognitive tool to learn to think mathematically; with them the users can operate not only with numbers, but also with symbols. Computers are machines for storing and dynamically manipulating symbols; they are capable of real-time programmable interactions with users. In this research work we visualize the computer as a cognitive tool.

Pea (1987) defines a heuristic taxonomy of certain transcendent functions that must be included in specific software, so that the computer can work as a true cognitive tool and promote students' cognitive activities in the mathematical learning process; he identifies two types of functions: purpose functions and process functions.

- Purpose functions:
  The key idea of these functions is that software promotes students to be part of what they learn and not to just limit them to be executors of instructions. That is, software must give users the opportunity to generate parts of the problem resolution process or exploration of the concepts.

- Process functions (with three aspects)
  (a) Tools to develop conceptual fluency.
  Software supports the students in routine and laborious tasks, allowing them to focus on cognitive tasks of higher levels and to save efforts in problem solving.
  (b) Mathematical exploration tools.
  Tools support students to explore concepts, recognize patterns and mathematics systems properties. Using these programs, students pose and test conjectures, and may discover theorems concepts and properties of mathematical concepts.
  (c) Representation tools.
  Tools enable students to relate different representations from the same concept and relate it to other concepts. This way, the concept in question may be seen from different angles.

In Fathom we observe all the previous functions; therefore we consider it as an important cognitive tool for learning probabilistic and statistical concepts.

Dörfler (1993) based on Pea´s ideas, identifies different ways in which integrating computer tools into mathematics teaching can help students to reorganize their cognitive systems:

1. *Change the activities to higher cognitive levels (meta-level).*

Computers may support actions of higher cognitive level by means of summarizing and simplifying complex processes and by entities, which are easily manipulable. For this to happen, a deep knowledge and experience about the way the tool works is necessary. The following examples of meta-level activities are worth to be mentioned: writing of computer programs, thinking about diverse phases of calculation that a problem requires in relation to the same calculations that a computer realizes in an automatic manner, selecting scales for a graph to adapt it to specific purposes, etc.

2. *Change of objects with which activities are carried out.*

The use of technological tools brings with it a change in the objects with which they work. Therefore, technological tools do not only change the structure and the form of an activity, but also its content. For example, when we use a statistics software, the set of objects are extended, and so tables with data of the samples and populations, tables with descriptive statistic, as well as formulas and different types of graphics are considered as part of the same analysis. These representations may become objects of mental activities, when changing values of parameters, data and scales to see their effect to other objects with which they are related. This capacity of the software entails a reorganization of the mental activity and a change in the approach at a higher cognitive level.

3. *Focus the activities in transformations and analysis of representations.*

Processes that involve problem solving and other cognitive processes often can be guided and organized in a successful way by concrete representations, images or models of the particular situation. Then, the mental processes essentially consist of transformations and manipulations of these representations. To support this process the computer offers a great variety of graphical, numerical and symbolic elements for the construction and manipulation of representations. Therefore, the user can construct diverse representations of many situations and analyze them on the computer screen.

4. *Support the situated cognition and problem solving*

The situated cognition theory postulates that genuine learning is reached in the investigation of the qualities, relations and elements of situations in which the student is involved. So, the computer can support to solve problems through its facet of simulation. Computers can help the students construct a bridge between statistics and reality allowing the access to modeling of concrete situations and the real data.

Methodology

The research was carried out with a group of 15 volunteer students, who took a probability and statistics course in the Computer Science Faculty of the Autonomous University of Sinaloa (Mexico). Their background in probability is the same as that of the majority of the students who are enrolled in the university, because high school courses are mainly focused on descriptive statistics and combinatorial techniques to calculate probabilities. Students already learned some Fathom commands related to exploratory data analysis in earlier study phases; there were two specific familiarization sessions in our study to learn the commands for working with probability distributions, especially the *randombinomial*, *randomnormal* and *randompick* commands.

In the selection and design of the activities we decided that the students should utilize probability tables or distribution formulas as well as simulation to calculate probabilities. In this last case, we explored the effect of the number of simulations on the precision of the results, the match between theoretical and empirical distributions. Besides, we considered extreme values for the variable that usually do not appear in probability tables, with the intention of investigating student's ideas about the impossibility or small possibility for such values to occur in reality.

In total, four activities were designed; two corresponding to the binomial distribution and two to the normal distribution (see Appendix). In each session, a worksheet with instructions to answer the questions using probability tables and software was given to the students, who saved their daily work in computer archives. The work of some students was videotaped, and at the end some of them were interviewed about issues that were considered important by the researcher. The interviews followed a semi-structured format, in which the researcher defined certain points of interest to question the students, but posing new questions throughout the interviews, whenever it was necessary. In the section of analysis and discussion of results some significant fragments of the interviews are described.

Results and Discussion

In the Fathom environment the students had access to symbolic, graphical and numerical representations to model the distributions, generate results and calculate the required probabilities. These representations are linked in an automatic way by the software and are visualized on the computer screen in a simultaneous way, which allows the students to recognize and establish relations between diverse concepts involved in probability distributions. For example, how the number of simulations influences the quality of approximation of the theoretical probability distribution by the frequency distribution of the results (i.e. the empirical probability distribution); to visualize the probabilities for each value of the random variable by means of graphics, how specific values of a parameter influence the shape of the distribution (see Figure 1). The following is an example of the first activity for the binomial distribution:

The Telektronic Company buys enormous lots (with thousands) of fluorescent light bulbs and uses the following acceptance sampling plan: They take a random sample of 20 light bulbs and test them; the lot is accepted only in case one find only 0 or 1 light bulb to be defective in the sample. One particular lot of thousands of light bulbs has in reality (let us assume that for the while) a defective rate of 5%.

a) Identify and define the random variable of interest.
b) Determine the probability distribution of this variable.
c) What is the probability that the lot is accepted if the described acceptance procedure is applied?
d) Compare the results obtained by probability tables to results obtained by simulation.

Photo 1. Students during one work session of the study.

At the beginning of each (binomial) problem the students were asked to define the random variable of interest by their own word, instead of denoting it in generic form by means of a symbolic X. The purpose of that approach was that in this way they related the information provided in the problem (parameters of the distribution) to the questions, which they had to answer; in addition, we considered that the verbal definition of the random variable in the context of the problem could be useful to identify with more facility the possible values that the random variable could take in the sample.

However, we observed that the students had difficulties to identify the random variable and define it in their own words; these students began the simulation defining the variable only in a generic form by means of X. For example, in the light bulb problem (problem 1, see Appendix), they incorrectly used phrases like “probability to accept the lot”, “probability that they are defective”, “number of times that the lot was accepted", when they should use “the quantity of defected light bulbs in the sample”. Despite the previous difficulties, these students identified the parameters to simulate the distribution.

The tasks related to the binomial distribution (problem 1 and 2; see Appendix) were solved correctly by 12 and 13 students respectively using probability tables, while all 15 students were successful in the Fathom environment. The tasks related to the normal distribution (problem 3 and 4; see Appendix) were solved correctly by 14 students in both environments, but with little discussion about the results obtained, particularly when they compared the empirical results to the theoretical distribution.

The methodology to address the tasks in the Fathom environment involved starting with simulation of a few samples, superposing the theoretical distribution to the results (i.e. the empirical frequency distribution) and continuing to simulate more samples until a sufficient match of empirical and theoretical distribution was reached at. This match sometimes is obtained with one or two hundred simulations, but in some cases one or two thousand simulations are required to have a satisfactory match, due to the fluctuations of random phenomena.

The purpose to superpose both distributions right from the beginning of the simulations was for the students to establish connections between both representations. This could enhance their awareness that the number of simulations has an impact on the precision of estimating the required probabilities by the relative frequencies, and to identify possible errors in the expressions to generate both distributions.

Not all the students selected the same number of samples in the simulation. For example, below we describe the solution of a student (Francisco). Once he established the model, he simulated 20 samples of size 20. Then he superimposed the theoretical distribution of probabilities to the empirical results obtained by simulation (histogram). When he observed that there was not sufficient correspondence between both distributions, he generated more samples until he had accumulated 2000 samples, to obtain an almost perfect adjustment with the theoretical distribution (see Figure 2).

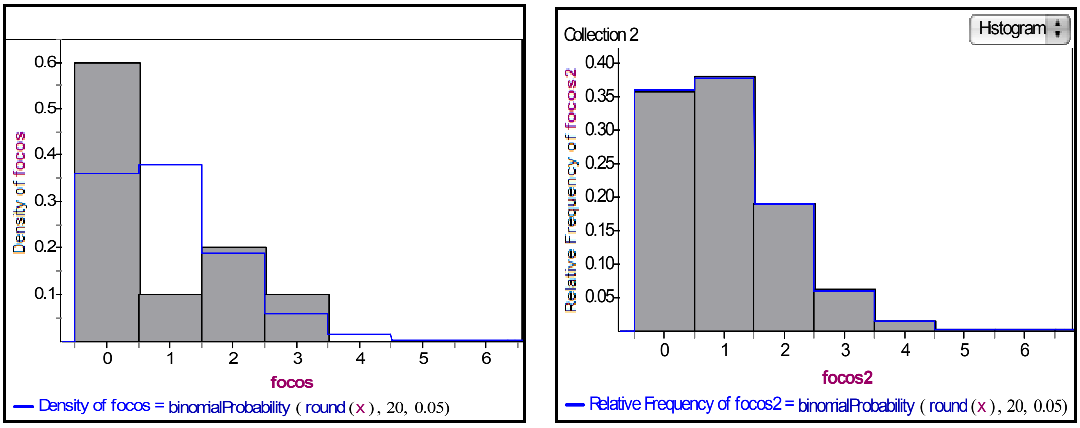


Figure 2. Empirical and theoretical distributions for 20 y 2000 samples respectively.

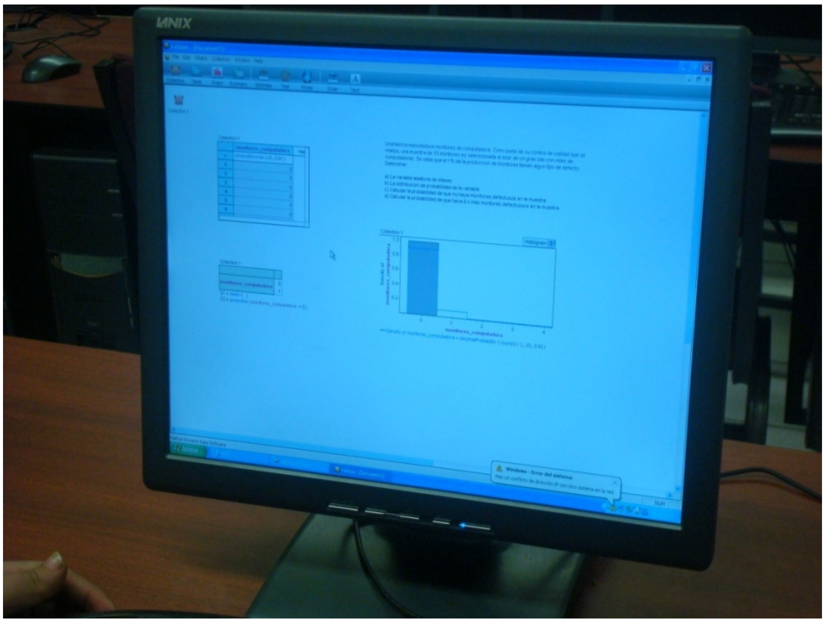
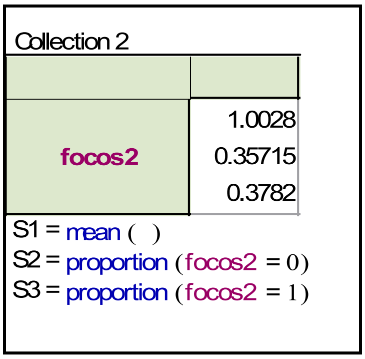


Photo 2. Running more simulations to fit the empirical distribution to the theoretical distribution in activity 2.

In addition, the students added tabular and symbolic representations (summary tables) to the previous graphs. Francisco simulated 2,000 samples in the light bulbs problem (see Figure 3) in order to estimate the probability that the sample is accepted.



Collection 2

focos2	1.0028 0.35715 0.3782

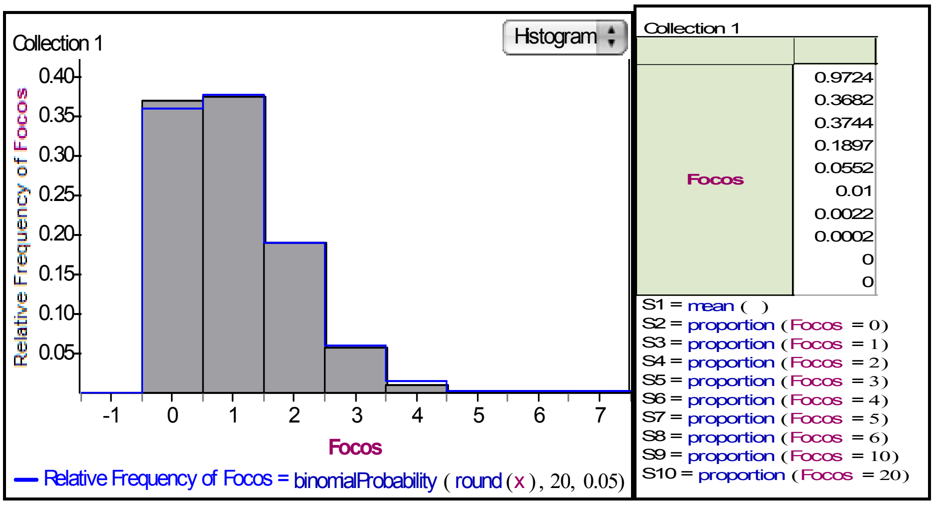
S1 = mean ( )
S2 = proportion (focos2 = 0)
S3 = proportion (focos2 = 1)

Figure 3. Summary table with proportions of cases for 2,000 samples.

In the same problem, another student (Ana) took 10,000 samples and constructed a distribution with a greater number of values (see Figure 3). Like Francisco, other students began generating a quantity of samples equal to the size of the sample in question. This means, they confused sample size with number of samples. This confusion only appears in the problems of binomial distribution, where one of the parameters is the sample size and where it is required to carry out the simulation. Similar difficulties have been found by Inzunsa and Sanchez (2005) in other studies about sampling distributions in a simulation environment.

The students, who incurred in the above mentioned error, noted that there was a noticeable difference between the graph of sample results and the theoretical distribution -or they noticed a discrepancy to the results of probability tables-, which forced them to review their simulation. This way, the complement between the diverse representations that the students put into play, and particularly the dynamic and visual characteristics of Fathom served like control resources that guided the students in the resolution of the problem.

The work developed by Ana (see Figure 4) reveals that she was aware of the values that the random variable “number of defective bulbs in the sample” could take, because she calculated the probability for various values as shown in the next table. Ana did not calculate the probability for all values, since she was aware that the probability was almost zero for many values, but calculated the probability for 20 defective bulbs, which was the maximum value that the variable could take.


Collection 1
Histogram
Relative Frequency of Focos
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
-1
0
1
2
3
4
5
6
7
Focos
Relative Frequency of Focos = binomialProbability (round (x), 20, 0.05)

Collection 1

Focos	0.9724 0.3682 0.3744 0.1897 0.0552 0.01 0.0022 0.0002 0 0

S1 = mean ( )
S2 = proportion (Focos = 0)
S3 = proportion (Focos = 1)
S4 = proportion (Focos = 2)
S5 = proportion (Focos = 3)
S6 = proportion (Focos = 4)
S7 = proportion (Focos = 5)
S8 = proportion (Focos = 6)
S9 = proportion (Focos = 10)
S10 = proportion (Focos = 20)

Figure 4. Graph with theoretical and empirical distribution and a summary table with probabilities of some values of the variable.

Some probability tables report a probability equal to zero (with three or four decimals) for values with low probability of a random variable; other tables left the respective space blank. This brings as a consequence that many students develop some misconceptions about the low probability values and consider these cases like impossible to happen. In our study we address this situation by asking the students to establish connections between the results that are provided by the tables and formulas with those that are provided by the simulation of a great amount of samples.

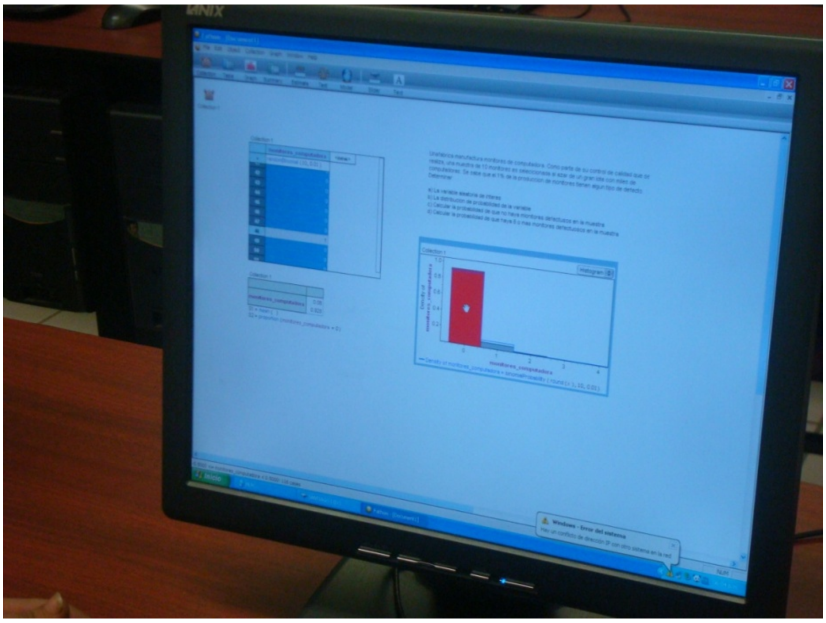


Photo 3. Selecting the value to 0 computer screen defectives in activity 2 (binomial distribution).

An important aspect that we explored in both types of distributions were values that are less probable and therefore do not appear in probability tables, so that the students became aware of the limitations of tables. For example, in the second problem for the binomial distribution (see the Appendix) we asked the probability of having 8 or more defective computer screens from factory production, in a sample of 10 screens with a 1% proportion of defects. The theoretical probability of this value is 4.44E-15, an extremely small probability.

After consulting the tables and even generating a simulation of 10,000 samples again (with no such case to happen), some students considered that given that probability in the tables was equal to zero, the event was impossible to happen. However, the calculus with the formula made them reflect that in reality the probability was very small, but not equal to zero. A simulation of 100,000 samples did not show any value, however, the students were conscious that a bigger number of simulations would be necessary to find some values, given the small probability.

Let us see Ana's case (A) who in the context of this activity had an interview with the researcher:

R: What is the probability of 8 or more defectives screens?

A: Zero, it is impossible

R: If we produce 1000 samples what will happen?

A: Still looks like zero

R: Look at the calculation with the formula to see if the results coincide

A: It is a very small number; almost zero.

R: But it is not zero, what does it mean?

A: It is less probable to occur

R: Observe that the result has 15 zeros after the decimal dot; and we have 28,000 samples. Do you believe that you can find any sample with 8 or more defective screens?

A: I would say no; even if the formula tells me the opposite

R: Do you have an idea of the quantity of samples that we should take in order to get at least a favorable case?

A: I would have to simulate millions of samples, then, maybe, one will show

We observe that Ana has difficulties to differentiate between impossible and very improbable events. On the one hand, the binomial formula shows results different from zero, and on the other hand, in 28,000 samples simulated with Fathom, no single positive result appeared. However, Ana is conscious that when the probability is so small, a lot of simulations are required to be able to find a favorable result.

An example of the activities done with the normal distribution in the study is shown next. Pregnancy time is (assumed to be) normally distributed with a mean of 268 and a standard deviation of 15 days. A woman says she will not give birth before the 308th day. Will that prediction become true? What is the probability that this will actually happen?

The formulation of the model of the normal distribution was simple for the students. Differences were observed in the quantity of data generated, which varied from 1000 to 10,000.

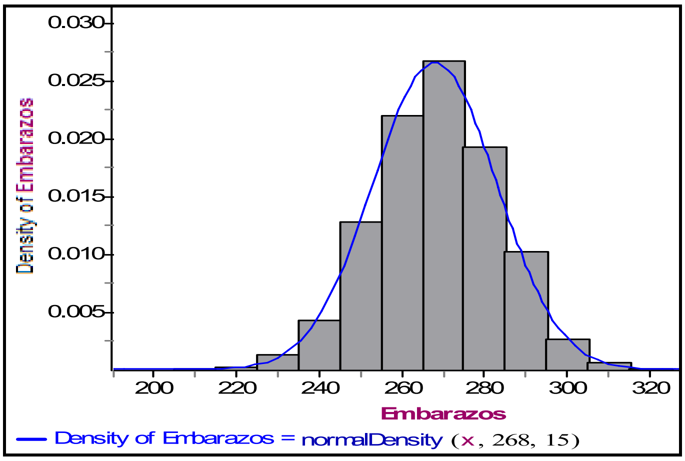


Figure 5. Graph with empirical and theoretical normal distribution.

For example, some answers that the students gave to this problem are as follows:

- It is very rare that a woman gives birth after 308 days; the probability is just a little greater than 0.004, but this is not impossible (Saul).
- The woman could be right in her prediction because the probability is a little greater than 0.0038. The probability is small, but could happen if you have many cases (Daniel).

Conclusions

In this study we explored an alternative approach to teaching probability distributions, based on the technique of computer simulation. We have considered important aspects like:

- The connection between empirical results provided by simulations and theoretical results provided by formulas and probability tables.
- Estimating probabilities by simulation and its contrast to theoretical probabilities, with special interest on small probability values that do not appear in probability tables.
- The effect of the number of simulations in the approximation of empirical results with theoretical results.

Our study shows that the simulation of random phenomena involved in the proposed problems is a simple activity for the students, due to the cognitive characteristics, user-friendly features, and interactive facilities of Fathom. In addition, the activities developed by the students in solving problems were substantially different in cognitive terms with respect to the one required in the traditional environment based on formulas and probability tables.

With Fathom he students have to undergo different clear-cut stages of the complete resolution process, which go from the selection of appropriate commands with parameters of the distribution, the simulation of cases to obtain results, the construction of graphs to display them and the use of summary tables to calculate probabilities of certain values or ranges of values of the variable. In the traditional environment this activity consists in substituting the values of the parameters and the value of interest of the random variable in a formula, or entering a probability table with the parameters of the probability distribution – in the case of the normal distribution with standardized Z-values.

The connection between empirical and theoretical results was an important aspect that was emphasized throughout the study. This way, the students were able to see that a frequency distribution of the simulation results comes closer and closer to the theoretical probability distribution as the number of simulations is increased; nevertheless when this did not happen, the results of probability tables and formulas were a reference to check that a parameter of the distribution was incorrectly chosen.

Such was the case of students who confused the sample size with the number of samples to simulate -in the case of the binomial distribution-. This also happened with the values that do not appear in the probability tables which some students considered as impossible to happen. The students could see that the formula showed a very small probability for a value of a variable and through simulations they could verify that indeed it was in cases that occurred with small frequency, as noted by the comments of Saul and Daniel in problem 3 (see Appendix). However, there were other problems with very small probability values; there were no favorable cases, despite the large number of simulations performed. However the students were aware that it might happen as they increase the number of simulations (see the case of Ana).

A very important resource provided by Fathom to achieve this, was the superposition of the frequency distribution with the theoretical distribution in the same graph, and mainly with the possibility to run simulations repeatedly and to see how the fit becomes in general better the more samples are taken. An example of this is the work developed by Francisco. This undoubtedly represents an activity in the meta-cognitive level of a cognitive tool.

In relation to the previous paragraph, the graph with both distributions and a summary table with the frequencies of certain values of the random variable, made it possible to see in a global form the probability distribution, with which students could identify values with greater or lesser probability to occur, and to recognize that some unlikely values do not happen also with a great amount of simulations (see Figure 4, built by Ana). This generally is not emphasized in the traditional environment, which usually focuses on the probability calculus of isolated values, precisely because of the difficulties involved in calculating and graphing them.

Finally, the objects provided by the computer environment to carry out the activities are different to those from the traditional environment based on formulas and tables. In Fathom, the students used expressions to simulate cases (based on the random command); they used tables of generated cases and graphics to represent them, and summary tables for the calculation of frequencies (empirical probabilities). Such laborious (and not fully understood) processes like the standardization and calculation of values Z, that are necessary to calculate probabilities of a normal distribution are no more required in the computer environment.

Despite the limitations of the study, we believe that the results show that students can develop a more complete and better understanding of probability distributions when their study relates theoretical and empirical aspects through the use of computational environments. In this regard the suggestion of Fischbein and Gazit (1984) to promote the connection between empirical and theoretical results in the teaching of probability becomes important. We believe that computational environments are a great step forward into this direction.

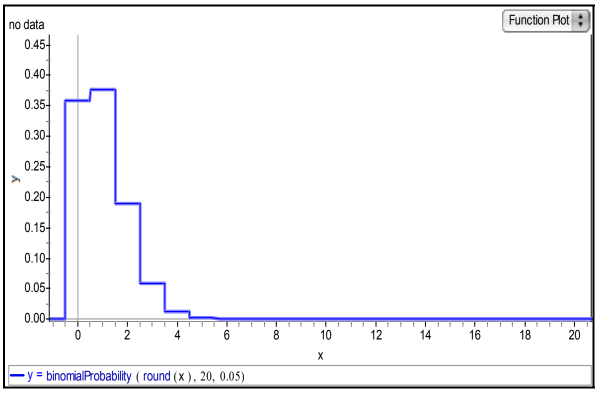
Appendix

1. The Telektronic Company buys enormous lots (with thousands) of fluorescent light bulbs and uses the following acceptance sampling plan: They make a random selection of 20 light bulbs and test them; the lot is accepted only if 0 or 1 light bulb is defective. One particular lot of thousands of light bulbs has in reality a defective rate of 5%.

a) Identify and define the random variable of interest.
b) Determine the probability distribution of this variable.
c) What is the probability that the lot is accepted?
d) Compare the results obtained with probability tables with results obtained *by simulation.*

The purpose of this activity is to find out if the students correctly identify and define the random variable associated with the information and questions of the problem (number of defective light bulbs in the sample). We also ask the students to construct the complete probability distribution and not only to calculate the acceptance probability of the lot, with the intention to have a more global idea about the potential values and their probabilities. In turn, we are interested in their comparison of results of probability tables (theoretical) with those obtained through simulation (empirical) so that they relate the effect of the number of simulations (observations) in the accuracy of results.

The theoretical probability distribution is shown in the following graph, built by Fathom.



The theoretical probability that the lots are accepted is determined by:

$$P(X=x)=\binom{n}{x}p^{x}(1-p)^{n-x} \quad x=0,1,\ldots,20$$

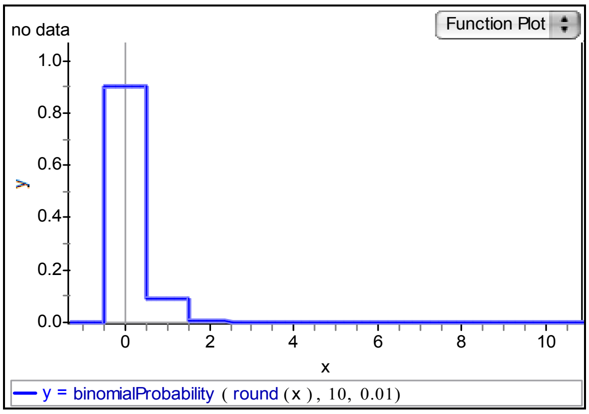
$$P(X=0)+P(X=1)=0.3584+0.3773=0.7358$$

2. A company manufactures computer screens. As part of quality control that is carry out a sample of 10 screens is randomly selected from a lot of production with thousands of screens. It is known beforehand that 1% of the production has some type of defect. Determine:

a) The random variable of interest.
b) The probability distribution of the variable.
c) Calculate the probability of non-defective computer screens in the sample.
d) Calculate the probability that there are 8 or more defective computer screens in the sample.

Additionally to the purposes of activity 1, in this activity we propose the research of students´ reasoning about the small probability values that do not appear in the probability tables.

The theoretical probability distribution is shown in the following graph, built by Fathom.



The theoretical probability to questions c and d is shown below:

$$P(X)=\binom{n}{x}p^{x}(1-p)^{n-x} \quad \text{x}=0,1,\ \ldots,10$$

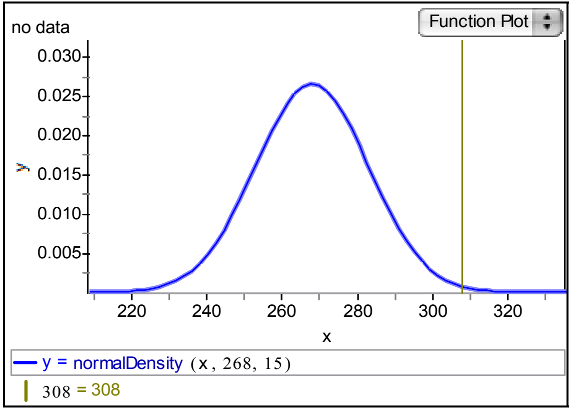
$$P(X=0)=0.9043$$

$$P(X\geq 8)=4.41E-15$$

3. Pregnancy time is (assumed to be) normally distributed with a mean of 268 and a standard deviation of 15 days. A woman says she will not give birth before the 308th day. Will that prediction become true? What is the probability that this will actually happen?

With this problem we propose to investigate students´ reasoning in connection to the normal distribution, when calculating probabilities of extreme values with very little probability; so that in a simulation of a relatively few cases this should usually not occur. The theoretical probability that birth occurs beyond 308 days of gestation of a baby under the conditions of the problem may be calculated to $P(X > 308) = 0.0038$.

The theoretical distribution and the indicated value are shown in the next graph:



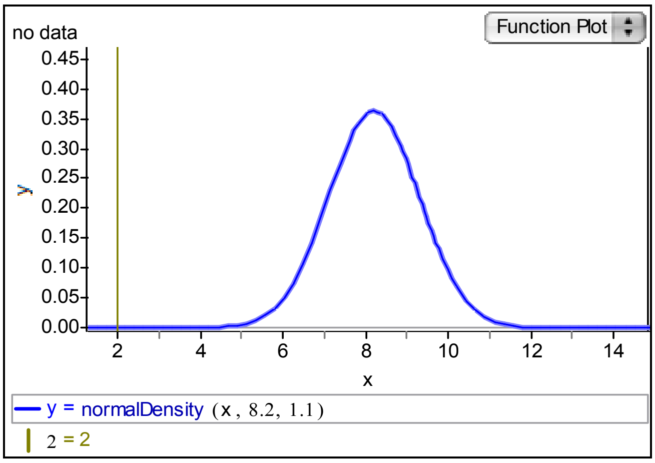
y = normalDensity (x, 268, 15)

308 = 308

4. It has been determined that the replacement time of a certain brand of television sets is normally distributed with a mean of 8.2 years and a standard deviation of 1.1 years. If the company offers a warranty of 2 years on their televisions, what proportion of television sets will be replaced?

Similarly, the problem involves random variable values that are very unlikely to happen, but in the context in which the problem is defined there is a great connotation for students, because practically the company would not have to replace any TV set given the problem´s conditions. The theoretical probability is given by $P(X < 2) = 8.68392E - 09$.

The theoretical distribution and the indicated value are shown in the next graph:



y = normalDensity (x, 8.2, 1.1)

2 = 2

References

Batanero, C., Henry, M. & Parzysz, B. (2005). The nature of chance and probability. In G. Jones (Ed.), *Exploring probability in school: Challenges for the teaching and learning,* New York, NY: Springer Verlag, 15-37.

Cohen, S. & Chechile, R. A. (1997). Probability Distributions, Assessment and Instructional Software: Lessons Learned from an Evaluation of Curricular Software. In I. Gal y J. B. Garfield (Ed.), *The Assesment Challenge in Statistics Education,* Voorburg Netherlands: International Statistics Institute, 253-262.

Dörfler, W. (1993). Computer Use and Views of the Mind. In C. Keitel & K. Ruthven (Ed.), *Learning from Computers: Mathematics Education and Technology,* Berlin, Germany: Springer Verlag, 159-186.

Finzer, W., Erickson, T. & Binker, J. (2002). *Fathom Dynamic Statistics Software*. Key Curriculum Press Technologies.

Fischbein, E. & Gazit (1984). Does the teaching of probability improve probabilistic intuitions? *Educational Studies in Mathematics,* 15, 1-24.

Inzunsa, S. & Sánchez, E. (2005). Effect of a computer simulation and dynamic statistics environment on the sampling distributions´ meanings. In G. M. Lloyd, M. Wilson, J.L. Wilkins & S.L. Behm, S. L. (Ed.), *Proceedings of the 27th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education.* Roanoke VA. PME-NA.

Inzunsa, S. & Quintero, G. (2007). The Information and Communication Technologies as Cognitive Tools in the Teaching and Learning of the Probability and Statistics. In A. Tremante, F. Welsch & F. Malpica. (Ed.), *Proceedings of the International Conference on Education and Information Systems, Technologies and Applications,* Orlando FL: International Institute of Informatics and Systemics, 141-146.

Jonassen, D. H. (1994). *Technology as cognitive tools: learners as designers*. http://itech1.coe.uga.edu/itforum/spaper1/paper1.html, Accessed October 20, 2007.

NCTM (2000). *Principles and Standards for Mathematics Education,* Reston VA: National Council of Teachers of Mathematics.

Pea, R. D. (1987). Cognitive Technologies for Mathematics Education. In A. H. Schoenfeld (Ed.), *Cognitive Science and Mathematics Education,* Hillsdale, NJ: Lawrence Erlbaum Associates Publishers, 89-122.

Pfannkuch, M. (2005). Probability and Statistical Inference: How Can Teachers Enable Learners to Make the Connection? In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning,* New York, NY: Springer, 171-189.

Ruiz, B., Huerta, A, & Batanero, C. (2006). An exploratory study of students´ difficulties with random variables. *Proceedings of Seventh International Congress for Teaching Statistics*. Salvador Bahía Brazil. ISI-IASE.

Tauber, L. & Sanchez, V. (2002). Introducing the normal distribution in data analysis course: specific meaning contributed by the use of computers. *Proceedings of Seventh International Congress for Teaching Statistics*. Salvador Bahía Brazil. ISI-IASE.

Wilensky, U. (1997). What is normal anyway? Therapy for epistemological anxiety. *Educational Studies in Mathematics*, 33, 171-202.