ICME 11 Mexico 2008
11th International Congress on Mathematical Education

Probability calculus and connections between empirical and theoretical distributions through computer simulation

Santiago Inzunsa
Universidad Autónoma de Sinaloa
México

In this article, we analyze results of a study with 15 university students about the way how probability distributions are constructed and how probabilities are calculated through computer simulation.

The connection to theoretical probability distributions, the probability of uncommon results and the effect of the number of simulation runs on the precision of the results are highlighted.

In the experiment we used the normal and binomial distribution. We chose Fathom as software for its cognitive potential

Antecedents

The conceptual step from data exploration to statistical inference has been considered as difficult for many students (Batanero, Tauber & Sanchez, 2001), due to the approach based on random variables defined on a probability space, requiring certain mathematical knowledge that many university students do not have.

In order to avoid such difficulties, it is common to use probability tables and formulas to calculate the requested probabilities, sometimes without students having any idea of where they come from and what their limitations are.

Since the emergence of diverse computer statistical packages, in particular, those that have been designed for statistical teaching purposes as it is the case with Fathom (Finzer et al. 2002), a different approach to the study of probability distributions is possible based on the simulation principle.

This approach allows for exploring concepts that underlie probability distributions, such as the effects of parameters, variation and the amount of repetitions of a random event to reach a specified precision.
Furthermore, it gives another additional type of representation to the unique numerical representations that are used in probability tables.

The connection between empirical and theoretical results for teaching purposes is advocated by various statistics educators and researchers.

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.

The Principles and Standards for School Mathematics (NCTM, 2000) for levels 9-12, recommends

emphasizing understanding of probability distributions and using simulations to construct empirical distributions.

Methodology

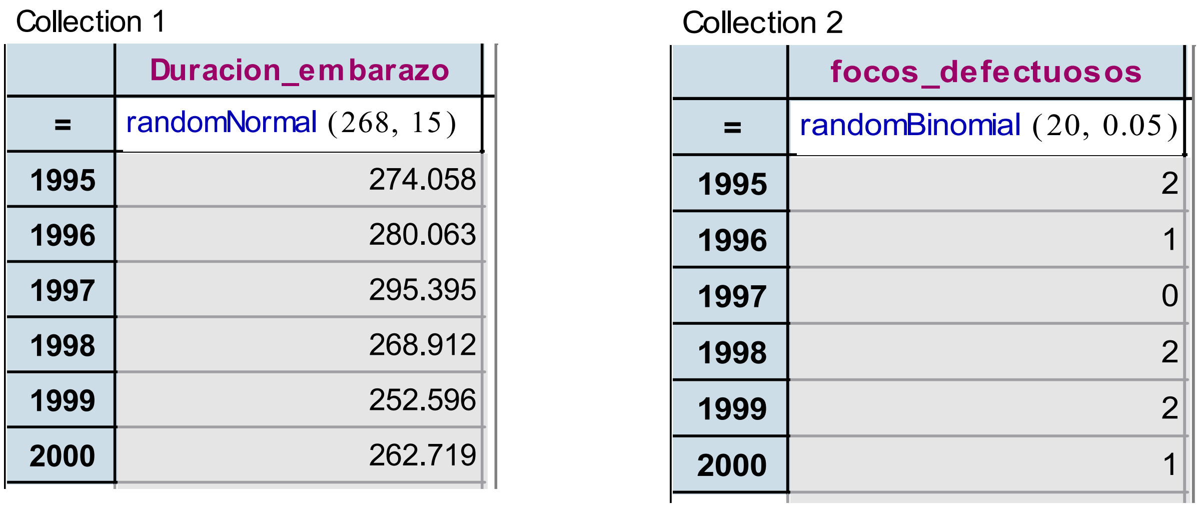
The research was carried out with a group of 15 students which were enrolled in a probability and statistics course in the Computer Science Faculty of Universidad Autónoma de Sinaloa (Mexico).

In the selection and design of the activities we decided that the students should utilize distribution tables or the distributions formulas, and simulation to calculate probabilities.

Four activities were designed, two of them corresponding to the binomial and two to the normal distribution. A learning process wa s designed where the participants themselves built a conceptual me aning based on the problems solved.

Results and discussion

The command *randombinomial (n,p)* generates in a direct form the values of the variable (success) in each sample of size *n* taken from a population with parameter *p.* For the normal distribution the *randomnormal* command is used (see Figure 1).



Collection 1

	Duracion_embarazo
=	randomNormal (268, 15)
1995	274.058
1996	280.063
1997	295.395
1998	268.912
1999	252.596
2000	262.719

Collection 2

	focos_defectuosos
=	randomBinomial (20, 0.05)
1995	2
1996	1
1997	0
1998	2
1999	2
2000	1

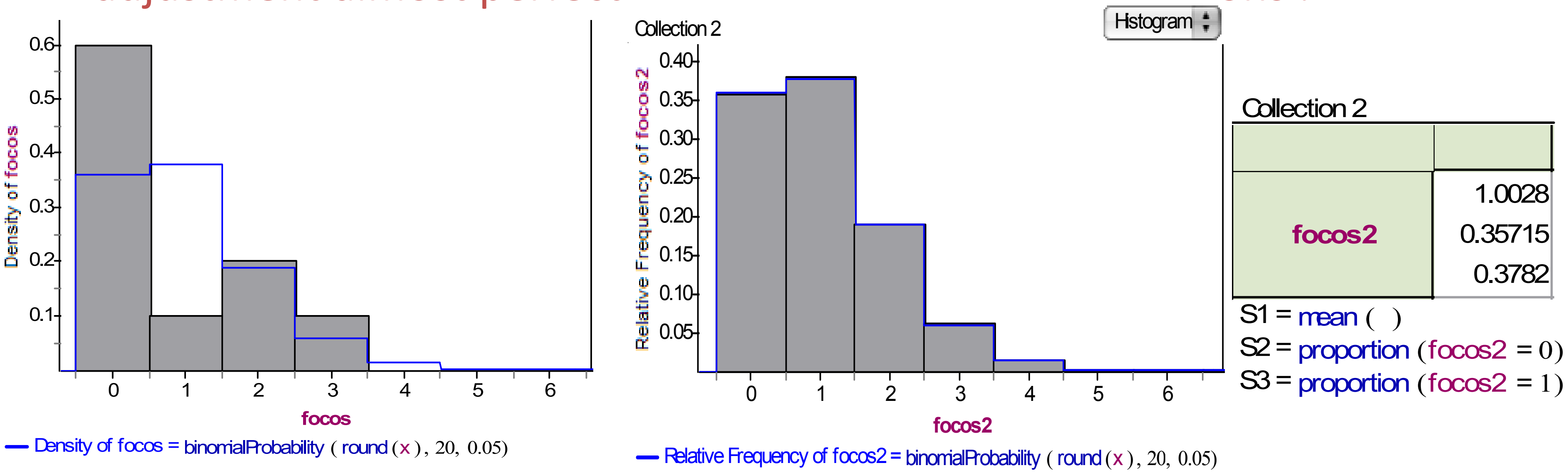
Fig. 1: Fathom´s case tables with 2000 results
generated for normal and binomial distributions respectively

The following is an example of the activities about binomial distributions.

The Telektronic Company buys big lots of fluorescent light bulbs and uses the following plan for acceptance sampling: They make a random selection of 20 light bulbs and test them; the lot is accepted only in case if 0 or 1 light bulbs are 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 to results obtained by simulation

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 simulated 20,000 more samples, to obtain an adjustment almost perfect with the theoretical distributions .



Collection 2

focos2	1.0028 0.35715 0.3782

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

Empirical and theoretical distributions for 20 and 20,000 samples respectively and summary table with proportions of cases for 20 000 samples

When consulting the distribution tables and even when generating a simulation of 10,000 samples, some students considered that if the probability of an event was equal to zero, this event was impossible to happen.

The calculations with the formula made them think that in reality the probability was very small but not equal to zero. A run for 100,000 samples showed no value, but the students were conscious that a generation of a greater number of simulations was needed to observe a value due to the small probability.

Conclusions

Due to the cognitive functions of Fathom, the students had a different approach to probability distributions respect to way they do it in a traditional environment by probability tables.

In the simulation environment, the students participated in the solution process, because the software is not a "black box" that shows the results by only introducing the necessary information.

The superposition of theoretical distributions to the empirical distribution, as well as the results obtained by probability tables and formulas, were a point of reference to evaluate the precision of the simulation.

With respect to that, the students point out that the results are very similar and that they are closer to the theoretical distributions with the number of observations or samples increased.

Besides allowing students to calculate the probabilities, the exploration and representation tools of the software enhances the solution, as they see the probabilities as proportions of cases in universe of results, in contrast to a probability table where the results represent (abstract) probabilities of a random variable in general.

In summary, the computer simulation allowed activities of higher cognitive levels, when modeling, exploring and calculating the probabilities.