STUDENTS' ERRORS AND DIFFICULTIES FOR SOLVING PROBLEMS OF SAMPLING DISTRIBUTIONS BY MEANS OF COMPUTER SIMULATION

Santiago Inzunza
Autonomous University of Sinaloa, Mexico
sinzunza@uas.uasnet.mx

This paper reports results the main errors and difficulties experienced by a group of eleven university students when solving problems of sampling distributions by means of computer simulation using Fathom *software (Finzer et al., 2002). The main difficulties were the formulation of the population model, the definition of the statistics to be calculated in each sample and the definition of the intervals to calculate the probabilities. It was not necessary to carry out some of the long processes of the pencil and paper environment which are the source of several mistakes and difficulties. It was also possible to facilitate the interpretation of some results such as the proportions of cases of interest out of the total of possible cases.*

INTRODUCTION:

In the literature of statistics education, (for example, Scheaffer, 1992; Moore, 1992) the advantages of exploring concepts related to the sampling distributions by means of computer simulation are highlighted frequently. Nevertheless, very little is said about the role that simulation plays in the solution of problems in which sampling distributions are involved.

Among the advantages of using simulation as a method to solve problems of probability and statistics, Biehler (1991) mentions:

1. The possibility of formulating models in concrete terms instead of expressing the ideas by means of symbolic models (representational aspect).
2. Students can process the data generated more easily than data generated using analytical and combinatory methods (computational aspect).
3. It is possible to begin with the design of the experimental environment instead of starting with the calculations (concept-model aspect).

According to Batanero *et al*. (2005), another advantage of simulation is the possibility to build pseudo-concrete models for many real situations and the possibility to work without a formal mathematical model, which allows one to act as an intermediary between the reality and the mathematical model.

The role of computer simulation as a method to solve problems, student's difficulties when using it, and the feasibility of using it as a tool for the student's arrival to acceptable solutions, compared to theoretical results, are just some of the questions considered as part of a research project about the meanings assigned, by university students, to the sampling distribution in an environment of computer simulation.

The problems used for this study are of the kind that requires deductive reasoning in order to be solved; this is to say that once the distribution of population and its parameters are known, what is looked for is the probability that some sampling results are obtained. This type of problem is very frequently used in textbooks and is studied before the study of statistical inference methods.

THE SOLUTION PROCESS

The concepts and actions needed to solve sampling distribution problems by means of computer simulation with Fathom have been identified into three well defined stages:

1. Formulate the population model.
2. Build the sampling distribution.
   a) Choose a sampling of a given size and define the statistic of interest in it.
   b) Repeat the process of selection of samples and make a collection with the statistic calculations in order to produce the sampling distribution.
   c) Divide the sampling distribution in parts in order to determine the proportion of statistics which are beyond a value or between two values given.

Most of the problems were solved in two ways: first by means of the computer simulation, and afterwards theoretically (using formulas and probability tables) so the students could make a comparison between the two results and give their opinion about the use of simulation as a method for solving problems.

1. *Formulation of the Population Model*

In order to formulate the population model for Fathom, it is necessary to identify the hypothetical elements of it, particularly the type of variable and the value of its parameters. In the case of discrete variables, the model is built by feeding the elements directly (box model). However, for the case of continuous variables, a formula which generates random numbers with the desired population characteristics is used.

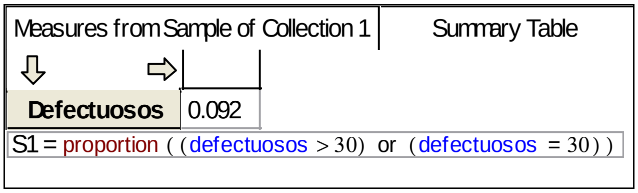
This stage was particularly difficult because the students did not take into consideration this difference. For most of the problems, two students at most were able to establish the model properly by themselves. For example, in one activity whose population was 30% of brown chocolates and the rest of another color (which was irrelevant for establishing the model), some students considered all the colors as equally probable, others gave 30% of probability to the brown ones and the rest assigned the colors with a random rate, but the correct thing to do was to consider only the brown ones and identify the rest of the chocolates with the same notation. Meanwhile, when they had to solve an activity with a continuous variable, the students neglected the type of variable and tried to establish the population as a discrete one.

2. *Construction of the Sampling Distribution*

Two stages were identified in the construction of the sampling distribution. In the first, many students had difficulties in writing the statistical formula, especially for the first activities. Meanwhile, for the second part of the process, the students made several mistakes because they took samples from the samples (double sampling) instead of taking population samples. Some students made an even more transcendental mistake: when asked to take one thousand samples of size ten, they took one sample of size thousand. This mistake is related to the difficulty of going from the results of a sample to a distribution. Saldanha and Thompson (2003) also reported this difficulty in a study with senior high school students and they defined it as a complicated stage in the process of acquiring the necessary skills to interpret the sampling distribution results.

3. *Division of the Sampling Distribution and Calculation of Probabilities*

In order to calculate the probability of a sampling result, the expression with the desired interval is introduced in a summary table (Figure 1). An additional resource is to use a formula to feed the sampling distribution table with the statistic values within the established interval. The advantage of this representation is that it allows shading the corresponding area in the graphic representing the sampling distribution so students can establish a relation between the darker area and the probability value obtained (Figure 2). Only two students (Mónica and Coral) used the two types of representations systematically when calculating the probability.



Measures from Sample of Collection 1	Summary Table
⇩ ⇨	
Defectuosos	0.092

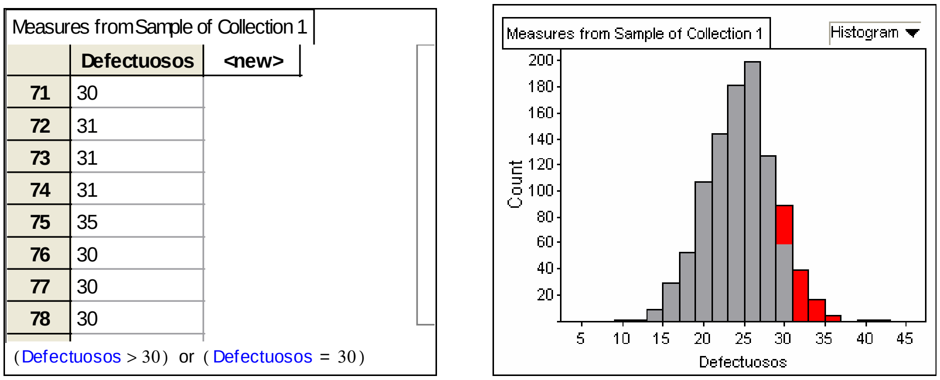
S1 = proportion ((defectuosos > 30) or (defectuosos = 30))

Figure 1: Summary table of the intervals of probability to be calculated

The main difficulties of this stage refer to the use of the connectors *and, or* and the inequality symbols (> <). This was the case for two students (Jorge and Omar) who in one of the activities used the connector *and* instead of *or* (Figure 3).

For another activity most of the students made the same mistake: they considered a single sample to calculate a specific probability instead of using the sampling distribution. In other words, they took the distribution of a sample as the sampling distribution. An example of this was taken from the teamwork of two students (Jorge and Gerardo) who took a sample of size forty and used the option *proportion* to calculate the probability that the mean will be smaller than 239 or bigger than 241, instead of using the sampling distribution (Figure 4). As a result of this confusion, the students misinterpret the simulation results as a percent of cases instead of a percentage of proportions or means of samples.

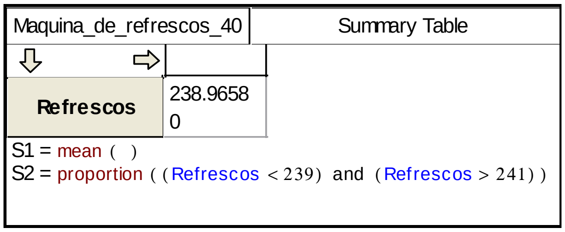
This mistake has also been reported by Saldanha and Thompson (2003) and Lipson (2002), who consider that its origin is the difficulty of the students to go from the results of a sample to the results of the sampling distribution.


Measures from Sample of Collection 1

	Defectuosos	<new>
71	30	
72	31	
73	31	
74	31	
75	35	
76	30	
77	30	
78	30	

(Defectuosos > 30) or (Defectuosos = 30)

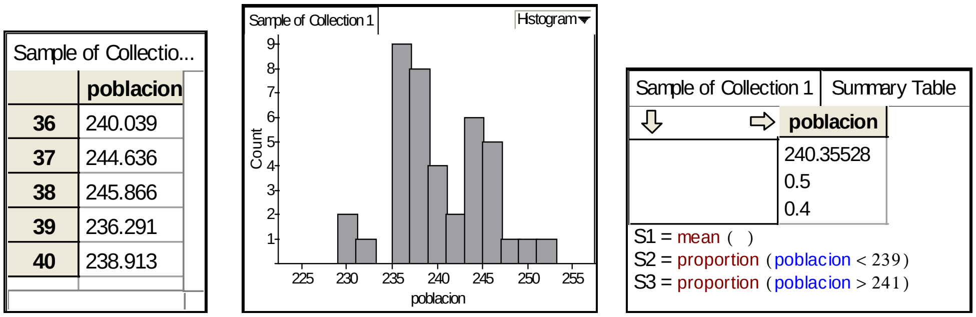
Figure 2: Representations used by Mónica and Coral to calculate probabilities


Maquina_de_refrescos_40 | Summary Table

Refrescos	238.9658 0

S1 = mean ( )
S2 = proportion ((Refrescos < 239) and (Refrescos > 241))

Figure 3: Jorge and Omar's mistakes when using connectors


Sample of Collectio...

	poblacion
36	240.039
37	244.636
38	245.866
39	236.291
40	238.913

Sample of Collection 1 | Summary Table

	poblacion
	240.35528 0.5 0.4

S1 = mean ( )
S2 = proportion (poblacion < 239)
S3 = proportion (poblacion > 241)

Figure 4: Confusion of sample distribution with sampling distribution

The mistakes and difficulties present in the simulation process were of different types and some of them were more persistent than others. In order to know which stage was the most difficult in the students' opinion, they were interviewed at the end of the study. The results are included in Table 1.

Table 1: Students' answers to the question: In your opinion, which stage of the process was the most difficult?

Student	*Answer*
Edgar	The definition of the population and the definition of the formula to collect the samples.
Omar	The interpretation of the expressions bigger or equal and smaller or equal to introduce the formula.
Jorge	Formulation of the population and the formula to calculate the statistic.
Denis	For me, it is confusing when I want to take some samples.
Donovan	The definition of the population and the calculation of the statistics.
Coral	When defining the population.
Mónica	When defining the population.
Viridiana	When defining the population.
Ana Lilia	The formulation of the population

From these data it is possible to say that for the students the most complicated stage was the formulation of the model, followed by the statistics definition and the construction of the intervals to calculate the probabilities.

CONCLUSIONS

Notwithstanding the mistakes and difficulties described above, the students were able to solve the problems of sampling distributions by means of computer simulation. They were helped by the researcher only when it was absolutely necessary, to enable them to continue working, and most of the help was for the formulation of the population model. The students' solutions were accepted as correct due to their closeness to the theoretical solution.

Among the advantages of the solution by means of simulation was the fact of interpreting the probability of the sampling results as proportions of cases of interest in a total number of possible cases, especially because of the characteristics of the software which allowed them to filter the results and even, in some cases, shade the graphic.

Moreover, with the solution of problems by means of simulation it was not necessary to carry out all the long and unavoidable processes of the pencil and paper environment, such as the standardization of the sampling distribution and the use of probability tables which are source of several mistakes when solving problems.

REFERENCES

Batanero, C., Henry, M. and Parzysz, B. (2005). The nature of chance and probability. In G. A. Jones (Ed.), *Exploring Probability in School: Challenges for Teaching and Learning.* (pp. 15-37). New York: Springer.

Biehler, R. (1991). Computers in probability education. In R. Kapadia and M. Borovcnik (Eds.), *Chance Encounters: Probability in Education. A Review of Research and Pedagogical Perspectives,* (pp. 169-212). Dordrecht: Kluwer.

Finzer, W., Erickson, T. and Binker, J. (2002). *Fathom Dynamic Statistics*^{TM} *Software*. Emeryville, CA: Key Curriculum Press.

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics,* Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Moore, D. S. (1992). Teaching statistics as a respectable subject. In F. Gordon and S. Gordon (Eds.), *Statistics for the Twenty-First Century*, MAA Notes #26, (pp. 14-25). Washington, DC: Mathematical Association of America.

Saldanha, L. A. and Thompson, P. W. (2003). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, 51, 257-270.

Scheaffer, R. L. (1992). Data, discernment and decisions: An empirical approach to introductory statistics. In F. Gordon and S. Gordon (Eds.), *Statistics for the Twenty-First Century*, MAA Notes #26, (pp. 69-82). Washington, DC: Mathematical Association of America.