"Quaderni di Ricerca in Didattica (Mathematics)", Numero speciale n. 7, 2020
G.R.I.M. (Departimento di Matematica e Informatica, University of Palermo, Italy)

Technology as a tool to understand sampling in binomial distributions

Nuria Begué[1], Carmen Batanero[2], María M. Gea[2] and Pablo Beltrán-Pellicer[1]

[1]Facultad de Educación, Universidad de Zaragoza, 50009 Zaragoza, Spain. [2]Facultad de Educación, Campus de Cartuja, 18071 Granada, Spain.

E-mail: batanero@ugr.es

Abstract. In this paper we describe some recurrent errors related to the work with sampling distributions, a topic which is compulsory in Spain for high school students. We additionally suggest how the sampling distribution for the mean and the range of samples taken from the binomial distribution, can be simulated using the software Fathom. Therefore, students can investigate these simulations with different values of the parameters for the binomial distribution and different sample sizes to understand properties of representativeness and variability of the sampling distribution and to discriminate the three types of distributions involved in sampling.

Résumé. Dans cet article, nous décrivons certaines erreurs récurrentes liées au travail sur les distributions d'échantillonnage, un sujet obligatoire en Espagne pour les lycéens. Nous suggérons en outre comment simuler la distribution d'échantillonnage pour la moyenne et la gamme d'échantillons prélevés à partir de la distribution binomiale, à l'aide du logiciel Fathom. Les élèves peuvent donc étudier ces simulations avec différentes valeurs des paramètres de la distribution binomiale et différentes tailles d'échantillons pour comprendre les propriétés de représentativité et de variabilité de la distribution d'échantillonnage et pour distinguer les trois types de distributions impliqués dans l'échantillonnage.

Subject classification numbers: 97D40, 97D70.

1. Introduction

Sampling is receiving increasing attention from statistics education research, given the relevance of sampling in simulation, which is the currently the recommended approach to improve the understanding of probability and statistical inference (Batanero & Borovcnik, 2016; Eichler & Vogel, 2014).

In Spain, the curricular guidelines, as well as the university-entrance tests for social-science high-school students (17-18-year-old), include sampling distributions. More specifically, according to the current guidelines (MECD, 2015), 14-15-year-old students are introduced to the notions of sample and population and to the relative frequency of an event and its convergence to its probability by using simulation or experiments. Moreover, in the second year of high school (17-18-year-olds), students learn the idea of sampling distribution and the difference between parameters and summary statistics.

Despite these guidelines, which are indeed similar in other countries, previous research reports that students do not perceive the essential properties of sampling distributions, possibly because the concepts involved in sampling require the idea of conditional probability, which is difficult for many students (Harradine, Batanero, & Rossman, 2011). The aim of this paper is to describe some of these difficulties and to suggest some activities that may help students overcome them.

2. Common errors in working with sampling distributions

The wide research on intuitive understanding of sampling started within the heuristics-and-biases programme by Kahneman and Tversky (as summarised in Kahneman, Slovic, & Tversky, 1982) where heuristics are conceived as unconscious actions guiding the resolution process of complex tasks. This research is summarized in Batanero and Borovcnik (2016). In particular, according to the representativeness, heuristic people consider just the similarity between the sample and the population in making a probabilistic judgment. An associated bias is the insensitivity to the sample size when judging the variability of the sample proportion. The recency heuristic gives a priority to the past sample results over the information on the population. In the gambler's fallacy, the subject believes that the result of a random experiment will affect the probability of future events. Positive recency happens when the subject assumes that the upcoming results will reproduce the observed pattern, and negative recency when the expectation is that the future results will compensate the observed results.

Harradine, Batanero, and Rossman (2011) suggest that students often confuse the three different types of distributions that intervene in sampling: a) the theoretical probability distribution that models the values of a random variable and depends on some parameter, such as the population proportion p in a binomial distribution; b) the distribution of sample collected from the population where we compute a statistical summary, e.g., the proportion of successes in the sample, which is used to estimate the population parameter p; c) the sampling distribution for the statistical summary (sample proportion in the example), that is, the probability distribution describing all possible values of the sample proportion in samples of the same size that are selected from the population.

Shaughnessy, Ciancetta, and Canada (2004) also described a series of studies directed to explore students' understanding of sampling variability. A prototypical task requires the students to predict the number of objects with specific properties when taking samples from a finite population and to explain their reasoning. A typical context is drawing from a container filled with 20 yellow, 50 red, and 40 blue candies. The students have to predict the number of red candies that will result in 5 consecutive samples of 10 candies (with replacement). The authors describe the following types of reasoning:

- *Idiosyncratic students* base their prediction of sample variability on irrelevant aspects of the task (e.g., preference, physical appearance, etc.).
- *Additive reasoning students* tend to predict the samples taking only absolute frequencies into account.
- *Proportional reasoners* use relative frequencies and connect proportions in the sample to proportions in the population; they tend to predict samples that mirror the proportion of colours in the container.
- *Distributional reasoning* involves connecting centres and spread when making the predictions of sampling variability. In addition to reproducing the proportion, their arguments also reflect random variation.

3. Using simulation to improve students' reasoning

A possible explanation for the above difficulties is the students' lack of experience with the process of sampling. Traditionally, statistical courses are based on solving textbooks problems, whose main purpose is getting competence with computing probabilities using the normal distribution table or, at best, the calculator. With this method the three distributions involved in the process of sampling may not be discriminated by the students.

A possibility to improve the teaching of sampling is asking the students first to solve a problem and later to simulate the situation using computer software. The interest of simulation in the teaching of

inference is highlighted, for example, by Eicher and Vogel (2014), and is supported today by the abundance of interactive resources on the Internet that allow to gain experience with repeated sampling and sampling distribution. Let's consider, for example, Task 1 (Figure 1):

Task 1. A parcel of 100 drawing pins is emptied out onto a table by a teacher. Some drawing pins landed "up" and some landed "down". The results were as follows: 68 landed up and 32 landed down .
The teacher then asked four students to repeat the experiment (with the same pack of drawing pins). Each student emptied the pack of 100 drawing pins and got some landing up and some landing down. In the following table, write one probable result for each student:

Daniel up: down:	Martin up: down:	Diana up: down:	Maria up: down:

Figure 1. Example task.

The mathematical model implicit in this situation is the binomial distribution with parameters n = 100 (sample size) and p (population proportion for the event in which we are interested). Since p is unknown, we estimate it by the proportion in a sample (which is an unbiased estimator of p with miminal variance; see Zacks, 2014). This sample was given in the text of Task 1. Since we ask for probable results, we expect that students will suggest values close to the distribution mean and then we can evaluate their perception of that mean. At the same time, the four values provided by each student will help us understand their perception of variability in sampling.
The first step is asking each student to complete their four predictions in a table. Some examples of responses by students to this task are presented in Table 1. These students are in the last year of high school (17-18-year-old students) and afterwards were asked to justify their responses, in which we can see a variety of reasonings. Then, while student S2 takes into account the frequentist probability that a pin would land up, according to the data in the task, students S1 and S4 provide values that do not consider these data and instead assume that both events in the experiment are equiprobable. Student S3 on turn, produces results in all the range of possible values in the binomial distribution and do not consider that the big number of experiments (n=100) correspond to a small variability around the expected value.

Table 1. Number of pins landing up in the predictions of some students.

Student	Daniel	Martín	Diana	María
S1	62	58	55	44
S2	72	70	65	67
S3	73	2	100	0
S4	43	38	53	48

Once the responses by the different students have been discussed, the teacher (or the students themselves if there are available computers in the classroom) can simulate the situation using some software, for example, Fathom. With this software it is very easy to produce, for example, 5000 simulations of the experiment consisting in taking four values at random from the binomial distribution $B(100, .68)$. When looking to the graphs of the distribution of the mean for these four values in the 5000 simulations (Figure 2), students easily can observe that the expected value (mean) is 68 and students who assumed the events were equiprobable can correct their reasoning.

Moreover, it is possible to plot some lines marking the different percentiles that limit the 68% and 95% central values in these distributions. The box plot also marks the median of the distribution (that coincides with the mean) and the lower and upper quartiles. In this way students can see that not all the values of the sample mean are equally likely, as the most probable values range around the theoretical expected value.

Figure 2. Box plot and histogram for the mean of four values in the B(100, .68)

The same conclusions may have been drawn theoretically, since the sample mean distributed as a normal distribution with expected value equal to the mean in the population (68) and standard deviation equal the standard deviation in the population divided by the square root of the number of experiments (10 in this case). Instead of working with the standard deviation which is more difficult to visualize, we can simulate the distribution of the range for the four values in the binomial distribution *B*(100, .68), which are represented in Figure 3 for 5000 simulations. This time the distribution is not symmetrical and only takes integer values. Again students can compare the most likely ranges in the simulated distribution of ranges with the ranges that resulted in their responses. The idea is making students observe that they attribute too much variability to the binomial distribution when the parameter *n* is big.

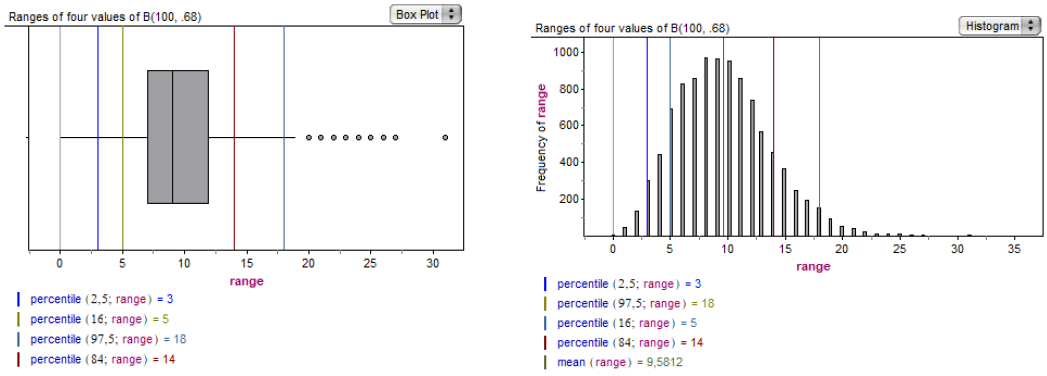
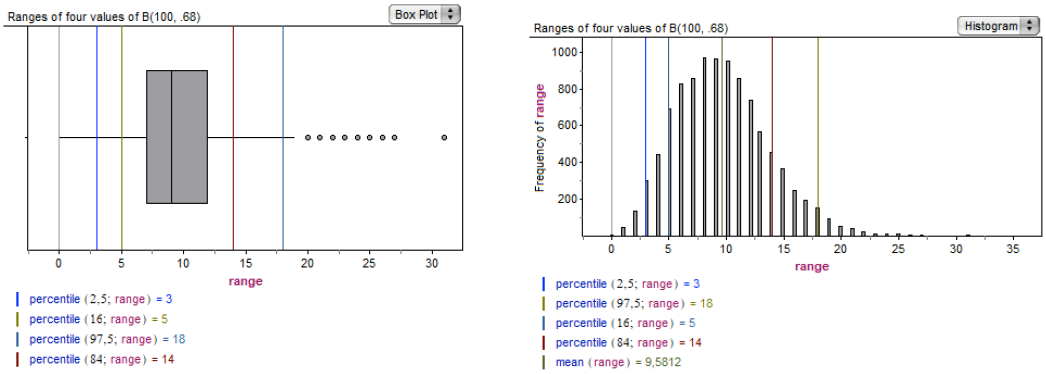


Figure 3. Box plot and bar graph for the range of four values in the B(100, .68)



Additionally, students can investigate with different values of the parameters for the binomial distribution and different samples sizes to understand properties of representativeness and variability of the sampling distribution and to discriminate the three types of distributions involved in sampling. As we did in the sample, students can obtain the central intervals containing 68% and 95% of the sampling distributions and compare with their predictions of probable values in samples of the binomial distribution. Finally, the simulation of the sampling distribution for the mean will be compared with the approximate distributions given by the Central Limit Theorem so that students can better understand when and why the substitution of the theoretical sampling distribution by the normal approximation is correct.

4. Final reflections

Sampling is an abstract topic. This is due in part to the calculation of the confidence interval using just a sample of the population. But inferential reasoning involves imagining every possible sample of the same size that could be taken from the given population. It is important to provide experiences involving simulations, such as the one described in the paper, so that students can observe that different sample from the sample population correspond to different values of the statistics (like the mean and the range), but that not every value of these statistics are equally likely.
Letting students observe the differences among the theoretical distribution (the binomial distribution in the example), the statistics in a particular sample (mean and range of the four values in the response of a student) and the sampling distribution of these statistics (the distribution of all the mean and ranges in 5000 different samples) would help them to clearly differentiate the three distributions implicit in the sampling. In this way they will become aware that the data distribution of the selected sample allows to make predictions about the probability distribution of the population. Nevertheless, the statistic of the available sample is only one element of the sample distribution of that statistic, which is the one that allows us to complete the margins of error and the related probabilities in the inferences about the parameter in the population.

Acknowledgements

Project EDU2016-74848-P (MEC), Group FQM126 (Junta de Andalucía) and Group S36_17D-Research in Mathematics Education (Government of Aragón and European Social Fund)

References

Batanero, C. & Borovcnik, M. (2016). *Statistics and probability in high school.* Rotterdam: Sense Publishers.

Eichler, A. & Vogel, M. (2014). Three approaches for modelling situations with randomness. In E. J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking, presenting plural perspectives* (pp. 75–99). Dordrecht: Springer.

Harradine, A., Batanero, C., & Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 235–246). New York: Springer.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

MECD (2015). *Real Decreto 1105/2014, de 26 de diciembre, por el que se establece el currículo básico de la Educación Secundaria Obligatoria y del Bachillerato*. Madrid: Autor.

Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2004). Types of student reasoning on sampling tasks. *Proceedings of the 28th Conference of the International. Group for the Psychology of Mathematics Education* (vol. 4, pp. 177-184).

Zacks, S. (2014). *Parametric statistical inference: Basic theory and modern approaches.* Oxford: Pergamon Press.