DO HIGH SCHOOL STUDENTS UNDERSTAND THE SAMPLING DISTRIBUTION OF A PROPORTION?

Nuria Begué[1], María M. Gea[1], Carmen Batanero[1], Pablo Beltrán-Pellicer[2]
[1]University of Granada (SPAIN)
[2]University of Zaragoza (SPAIN)
mmgea@ugr.es

The Spanish curricular guidelines, as well as the entrance to university tests for social science high school students (17-18-year-olds) include sampling distributions. To analyse the students' understanding of this concept we proposed to 127 students in Spain a questionnaire with four sampling tasks, where the sample size (n= 100 and n=10) and the population proportion (equal or different to .5) were systematically varied. The analysis of students' responses suggests a good understanding of the relationship between the theoretical proportion in the population and the sample proportion. However, the sampling variability was overestimated, particularly in big samples. We also observed the equiprobability and recency biases, as well as a deterministic conception in the students. The effect of the task variables on the students' responses is also discussed.

INTRODUCTION

The concepts involved in sampling are receiving increasing attention from statistics education research, since ideas linked to sampling underlie the work with simulation, which is the currently recommended approach to improve the understanding of probability and statistical inference (Eicher & Vogel, 2014; Huerta, 2018). Moreover, sampling establishes a bridge between statistics and probability and plays a key role in the study of topics such as the frequentist approach to probability or the law of large numbers.

In the Spanish curricular guidelines (MECD, 2015), the concepts of population and sample, as well as the frequentist approach to probability appear in the first two grades of secondary education (12-14-year-olds). More specifically, according to these guidelines, students in these grades 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. In the third grade (14-15 year-olds), students learn different methods of collecting samples, are introduced to the idea of representativeness, and are asked to judge the sample representativeness through analysis of the selection procedure, in simple cases. Finally, in the second year of high school (17-18 year-olds) students are introduced to the idea of sampling distribution and to the difference between parameter and summary statistics. They also intuitively use the central limit theorem to determine the sampling distribution for means and proportions.

Despite these guidelines, previous research carried out in other countries suggests that students do not perceive some properties of sampling distributions. A possible reason is that parts of the concepts involved in sampling involve the idea of conditional probability, which is difficult for many students (Borovcnik, 2012). In order to assure that these difficulties are not maintained in the Spanish students by the end of high school, this study focuses in analysing their understanding of the relationship between the population proportion, the expected value of a sample proportion, as well as its variability in different samples and the effect of sample size on such variability.

BACKGROUND

Understanding sampling requires linking two apparently opposed ideas: the sample representativeness and variability (Rubin, Bruce, & Tenney, 1991; Saldanha & Thompson, 2002). The sample representativeness implies that a random sample of adequate size will approximately reproduce the population characteristics, whereas the sample variability indicates that the composition of different samples. For example, although the proportions of a given event in different samples of the same size approximate the population proportion (representativeness), we can obtain different sample proportions (variability) in different samples.

Understanding sampling will also require the students' identification of three different types of distributions (Harradine, Batanero, & Rossman, 2011):

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

- The *theoretical probability distribution* that models the values of a random variable in a population and depends on some parameter. In our research, we considered a random dichotomous variable, and the parameter of interest is the population proportion *p* of elements sharing a given property.
- The *distribution of a sample data*, collected from the population, where we compute the proportion of successes $\hat{p}$, which is a summary statistic in the sample and is used to estimate the population parameter *p*.
- While the parameter *p* is an unknown constant, the value of $\hat{p}$ is a random variable that varies in the different samples. As such, it is characterised by a probability distribution describing all the possible values of $\hat{p}$ in the different samples of same size that can be selected from the population. This distribution is called the *sampling distribution* for the proportion.

Since there is a one-to-one correspondence between the number of successes in a sample of n elements the proportion of successes in the sample, our students will be asked to provide the expected values for the number of successes in the sample. The probabilistic model that applies to this variable is the binomial distribution *B(n, p),* where *p* is the population proportion and *n* the sample size.

Research background

The wide research on sampling started within the heuristics and biases programme (Kahneman, Slovic, & Tversky, 1982), where *heuristics* are understood as unconscious actions guiding the resolution of complex tasks. Such heuristics simplify probability problems, but often led to reasoning biases. In our research we ask the students for a probable value of the sample proportion in samples of different sizes. In such task the following heuristics may apply:

- The *representativeness heuristic* (Tversky & Kahneman, 1974) appear when a subject only considers the similarity between the sample and the population in making a probabilistic judgment. According to these authors people expect the essential characteristics of a random process to be represented not only globally in a sequence of results, but also locally in each part. An associate bias is the *insensitivity to sample size* when judging the variability of the sample proportion. In the *gambler fallacy* the subject believes that the result of a random experiment will affect the probability of future events. We speak of *positive recency* if 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.
- The *availability heuristic* consists in estimating the probability of an event basing only on the facility to find examples of similar situations. A related bias is the equiprobability one (Lecoutre, 1992), by which the results of any random phenomenon are considered to be equally likely.

METHOD

A total of 127 high school students (17/18-year-olds) from two different schools, one in Huesca and another in Zaragoza (Spain), in total 6 groups of students took part in the sample. These students had studied the curricular contents of sampling the previous years. These students were given a questionnaire including four tasks the first of which is reproduced in Figure 1.

In this task the students were asked to provide four probable values for the number of drawing pins landing up, when emptying a parcel of 100 drawing pins. 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 p by the proportion $\hat{p}$ =0.68 in a sample. In Table 1 we include the values of parameters for the all the tasks, as well as the intervals containing the 68% and 95% of sample means and the ranges considered optimum or acceptable. The context of the items was emptying 100 drawings pins on a table (Item 1), flipping 100 or 10 fair coins (items 2 and 3) and throwing 10 balls to a basketball goal (item 4).

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. Each student emptied a packet of 100 drawing pins and got some landing up and some landing down. In the following table, write probable results for each student:

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

Figure 1. Example of task given to the students.

Once the questionnaires were collected, we performed a statistical analysis of the four responses provided by each student in each item and compared the results by group. The average value of the four estimates provided by each students was used to evaluate his/her intuitive understanding of the relationship between the population and sample proportions, while the range of these four values served to assess their intuitive understanding of sampling variability. The number of X of successes in a binomial distribution, *B(n, p),* is a random variable with expected value *np* and standard deviation, $\sigma = \sqrt{np(1-p)}$. Therefore, we considered that the student had a good intuitive understanding of the sample proportion, if the mean value of his /her four estimates was close to *np* (theoretical mean). This average value was also compared with the intervals containing 68% or 95% in the theoretical sampling distribution (see Table 1).

Table 1. Summary of the tasks proposed to the students

	Task 1	Task 2	Task 3	Task 4
Sample size	100	100	10	10
Sample population	0.68	0.5	0.5	0.7
Expected number of successes	68	50	5	7
Standard deviation	σ =4,66	σ =5	σ =1.58	σ =1.45
Interval containing 68% of sample means	[63.3-72,7]	[45-55]	[3.4-6.6]	[5.5-8.4]
Interval containing 95% of sample means	[58.6-77,4]	[40-60]	[1.8-8.2]	[4.1-9.9]
Optimum range	[10-20]	[10-20]	[3-6]	[3-6]
Acceptable range	[21-30]	[21-30]	[7-9]	[7-9]

Since the interval $\mu \pm 2\sigma$ contains 95% of the observations in a normal distribution, the students' understanding of variability was considered adequate when the range of the four values provided by each student was included with the interval (2σ, 4σ) (Gómez, Batanero, & Contreras, 2014). If the range was included between 4 and 6 standard deviations (containing only 5% of observations in the normal distribution), it was considered high, but acceptable; if the range is higher, the variability of estimates was considered to be excessive and if the range was smaller than two standard deviations, we considered there was too much concentration in the data. These two last cases implied a poor understanding of the sampling distribution variability.

RESULTS

Understanding of the expected value

In Figures 1 we display the distribution of the averages in the four estimates provided by each student in the different tasks. These distributions suggest (in general) a good understanding of the relationship between the population and sample proportions by the participants, given the proximity between the theoretical proportion in the population and the average value of the distribution of all the students' responses in tasks 2 to 4 (See also Table 2). However, in task 1, there is a difference of 10 points between the theoretical value (68) and the mean of all the estimations provided by the students ($\bar{x}$ =57.9). The reason is that in this task many people considered both results to be equiprobable. There are also some atypical values, corresponding to students whose responses are very different from those in the group.

This information is expanded in Table 2 where we show the percentage of students according to different types of response in each task. From this table, it follows that most students provided estimates whose average fall either in the interval that theoretically contains 68% of values for the sampling proportion or in the interval that contains 95% of the values. Only a few students present negative or positive recency biases. We also observe a 16% of students providing average values close to 50 in task 1, where they assumed equiprobability of results, as they did not consider the frequentist information provided in the item.

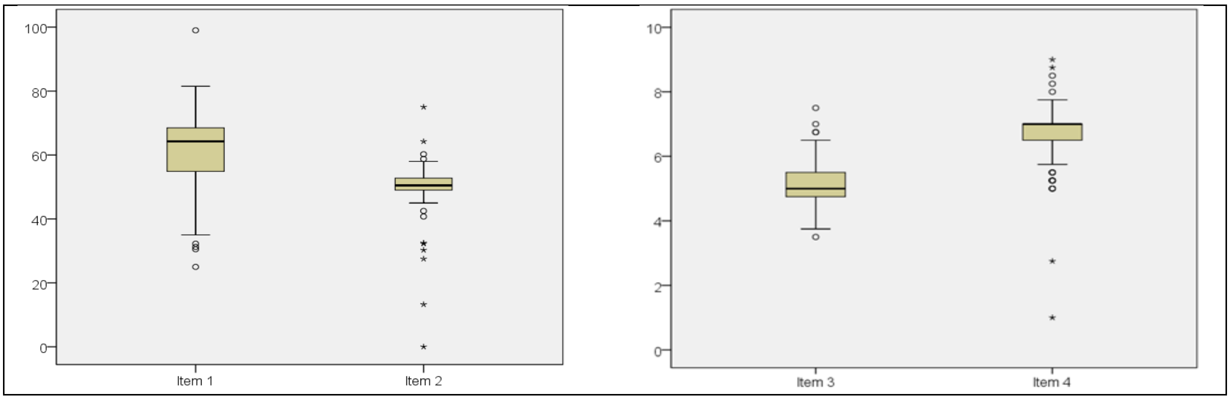

Figure 2. Distribution of average in students' estimates

Table 2. Percentage of students with average values in some intervals

Average value in Task 1 ($\bar{x}$ =57.9)	
Average value	%
Lower than45: Representativeness	9.4
45-55 (equiprobability)	16.6
63-73 (optimum)	65.4
Higher than 73	26.0

Average value in Task 2 ($\bar{x}$ =51.2)	
Average value	%
Lower than 45	7.1
45-55 (optimum)	80.3
Higher than 73	19.7

Average value in Task 3 ($\bar{x}$ =5.1)	
Average value	%
Smaller than 3.4	0
3.4-6.6 (optimum)	96.7
Higher than 6.6	3.3

Average value in Task 4 ($\bar{x}$ =6.6)	
Average value	%
Lower than 5.5	11.5
5.5-8.5 (optimum)	86.9
Higher than 8.5	1.6

Understanding the sampling variability

In Figure 3 we display the distribution of ranges for the four estimates provided by the student in each task. Comparing these distributions with the values established to analyse the ranges (Table 1), we conclude that an important percentage of students provide estimates with excessive variability in tasks 1 and 2 corresponding to big samples and most of them provided acceptable values of ranges in tasks 3 and 4 (small samples), contradicting studies such as those by Shaughnessy, Ciancetta, and Canada (2004). We also observed that 7 students provided four identical values in some question, four of them in all the questions, therefore denoting a deterministic conception of sampling.

In Table 3 we present the percentages of students providing estimates with ranges in different interval in each task. We observe an important percentage of students providing estimates with small variability (concentration) in the tasks corresponding to big samples (21.3% in task 1 and 29.9% in task 2). On the contrary the estimates for small samples had small or optimum variability. In particular in task 4 (throwing 10 balls to a basketball goal) 38.6% of students provided estimates very close to the theoretical value which translate in high concentration, since they interpreted the experiment in a deterministic way.

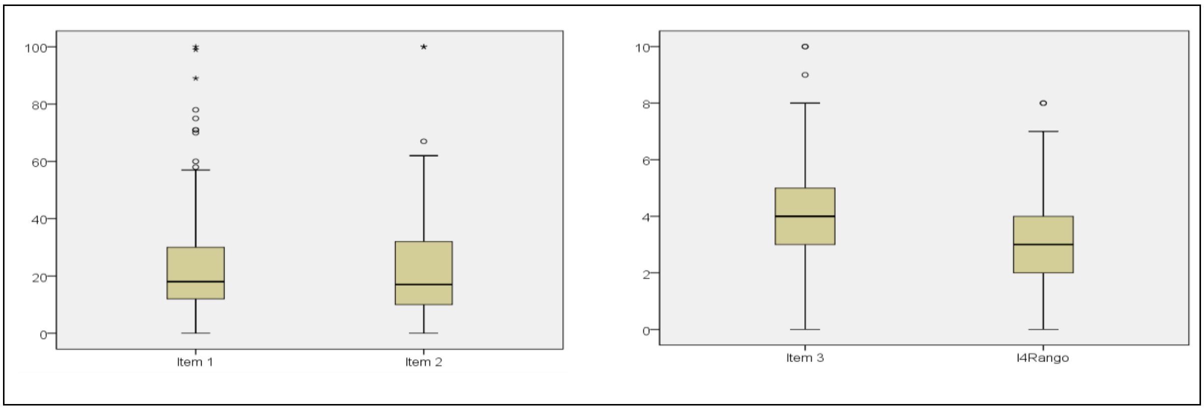

Figure 3. Distribution of ranges.

Table 3. Percentage of students with ranges in some intervals

Range	Task 1	Range	Task 2
Lower than 10 (concentration)	21.3	Lower than 10 (concentration)	29.9
10-20 (optimum)	38.5	10-20 (optimum)	29.2
20-30 (acceptable)	16.6	20-30 (acceptable)	17.3
Higher than 30. Excessive	23.6	Higher than 30. Excessive	23.6
Range	**Task 3**	**Range**	**Task 4**
Lower than 3 (concentration)	14.8	Lower than 3 (concentration)	38.5
3-6 (optimum)	77.8	3-6 (optimum)	45.8
7-9 (acceptable)	5.8	7-9 (acceptable)	5.7
Higher than 9.6. Excessive	1.6	Higher than 9. Excessive	0

Students' justifications

We additionally analysed the students' arguments when justifying their estimates in Item 1. These arguments were classified in the following categories: a) producing the estimates "at random", b) basing in physical properties of the random device, expressed in probability language or in everyday language; c) using an intuitive understanding of frequentist probability; and d) expressing ideas of variability. Given the lack of space, these arguments will be discussed in detail in the presentation at the conference.

CONCLUSION

The results suggest in general, a good perception of the expected value in the sampling distribution for proportions. Consequently, most students achieved a good level of proportional reasoning related to the sampling distribution (Shaughnessy et al., 2004). On the contrary, between 14.8% and 38.5% of participants (depending on the task) perceived correctly the effect of sample size on sample variability and therefore did not reach the distributional reasoning level sampling described by these authors.

These results suggest the need to improve the teaching of sampling and to provide students with some experience of the random variability within different samples of the same population. In this sense, the use of some simulation applets which are freely available on Internet can help the teacher to make students' conscious of sampling properties. Simulation, whose interest is highlighted by multiple authors (e.g. Huerta, 2018), is supported today by interactive resources that allow to gain experience with repeated sampling and sampling distribution. These resources are essential for the subsequent understanding of the inference since, although for example, the calculation of a confidence interval is performed from a single sample, inferential reasoning involves imagining every possible sample of the same size that could be taken from the given population (Saldanha & Thompson, 2002). The student would clearly differentiate the three distributions implicit in the sampling, becoming 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 sampling distribution of that statistic, which serve to complete the margins of error and the related probabilities in the inferences about the population parameters.

ACKNOWLEDGEMENTS

Project EDU2016-74848-P (AEI, FEDER); research groups FQM126 (Junta de Andalucía) and S119 (Gobierno de Aragón, FEDER).

REFERENCES

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

Borovcnik, M. (2012). Multiple perspectives on the concept of conditional probability. *Avances de Investigación en Educación Matemática, 2*, 5-27.

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. 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. 57-69). New York: Springer.

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

Gómez, E., Batanero, C., & Contreras, C. (2014). Conocimiento matemático de futuros profesores para la enseñanza de la probabilidad desde el enfoque frecuencial. *Bolema, 28* (48), 209-229.

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.

Huerta, M. P. (2018, in press). Preparing teachers for teaching probability through problem solving. In C. Batanero, & E. Chernoff (Eds.). *Teaching and learning stochastics: Advances in probability education research.* Hamburg: Springer.

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

Lecoutre, M. P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23*(6), 557-568.

Ministerio de Educación, Cultura y Deporte 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* (Royal decree establishing the structure of compulsory secondary-school and high-school curriculum). Madrid: Author.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics*. Dunedin, New Zealand: International Statistical Institute. iase-web.org/documents/papers/icots3/BOOK1/A9-4.pdf.

Saldanha. L., & Thompson, P. (2002) Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257-270.

Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2004). Types of student reasoning on sampling tasks. In M. Johnsen Høines, & A. B. Fuglestad (Eds.), *Proceedings of the 28th Annual Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 177-184). Bergen: PME group. Online: www.emis.de/proceedings/PME28/RR/RR045_Shaughnessy.pdf.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P.Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 117-128). New York: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974): Judgement under uncertainity: Heuristics and biases. *Science, 185*, 1124-1131.