WAYS IN WHICH HIGH-SCHOOL STUDENTS UNDERSTAND THE SAMPLING DISTRIBUTION FOR PROPORTIONS

CARMEN BATANERO
University of Granada (Spain)
batanero@ugr.es

NURIA BEGUÉ
University of Granada (Spain)
nbegue@correo.ugr.es

MANFRED BOROVCNIK
University of Klagenfurt (Austria)
manfred.borovcnik@aau.at

MARÍA M. GEA
University of Granada (Spain)
mmgea@ugr.es

ABSTRACT

In Spain, curricular guidelines as well as the university-entrance tests for social-science high-school students (17–18 years old) include sampling distributions. To analyse the understanding of this concept we investigated a sample of 234 students. We administered a questionnaire to them and ask half for justifications of their answers. The questionnaire consisted of four sampling tasks with two sample sizes (n = 100 and 10) and population proportions (equal or different to 0.5)systematically varied. The experiment gathered twofold data from the students simultaneously, namely about their perception of the mean and about their understanding of variation of the sampling distribution. The analysis of students' responses indicates a good understanding of the relationship between the theoretical proportion in the population and the sample proportion. Sampling variability, however, was overestimated in bigger samples. We also observed various types of biased thinking in the students: the equiprobability and recency biases, as well as deterministic pre-conceptions. The effect of the task variables on the students' responses is also discussed here.

Keywords: *Statistics education research; Sampling distribution; High-school students; Understanding of probability; Understanding of mean; Understanding of variability*

1. INTRODUCTION

The concepts involved in sampling are receiving increasing attention from statistics education research since ideas linked to sampling underlie work with simulation, which is the currently recommended approach to improve the understanding of probability and statistical inference (Batanero & Borovcnik, 2016; Eichler & Vogel, 2014; Huerta, 2018). Heitele (1975) included sampling in his list of fundamental stochastic ideas (see also Burrill and Biehler, 2011) as 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.

Our everyday knowledge is based on the idea of sampling, since we only can observe a small piece of reality. Moreover, results of surveys based on sampling play a crucial role in arguments in the media, so we need to help students develop their sense of sampling and make them conscious of reasoning biases that may afflict them. As a first step in improving the education around these ideas, the aim of this study is assessing high-school students' knowledge of sampling.

In the Spanish curricular guidelines (MECD, 2015), the concepts of population and sample as well as the frequentist approach to probability are covered in the first two grades of secondary education

(12–14 years old). 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 calculations, simulation, or experiments. In the third grade (14–15 years old), students learn different methods of collecting samples, they are introduced to the idea of representativeness, and, in simple cases, they are asked to judge the representativeness of samples through analysis of the selection procedure. Finally, in the second year of high school (17–18 years old), students learn the idea of sampling distribution and the difference between parameters and summary statistics. They also use the Central Limit Theorem intuitively to determine the sampling distribution for means and proportions. For all this content, the students are introduced to the sampling concepts and properties as well as to the required calculations to solve related problems.

Despite these guidelines and endeavours undertaken in teaching the concepts, previous research (as summarised in Castro Sotos et al., 2007 and Harradine et al., 2011) corroborates that students do not perceive the essential properties of sampling distributions correctly. A possible reason is that parts of the concepts involved in sampling require the idea of conditional probability, which is difficult for many students (Borovcnik, 2012). We base our study on a sample of Spanish students and investigate whether these difficulties with the concepts are resolved for the students by the end of high school.

The aim of this study is to analyse students' understanding of the relationship between the proportion in a population and the expected value of a sample proportion, as well as students' understanding of the variability of the value in different samples and the effect of sample size on this variability.

2. BACKGROUND

2.1. UNDERSTANDING THE SAMPLING DISTRIBUTION

Understanding sampling requires linking two apparently opposed ideas: the sample representativeness and sample variability (Chance et al., 2004; Rubin et al., 1991; Saldanha & Thompson, 2002; Watson & Moritz, 2000). Sample representativeness implies that a random sample of adequate size will reproduce the population characteristics approximately, whereas sample variability implies that different samples will be different. For example, although the proportions of a given event in different samples of the same size approximate the population proportion (representativeness), different sample proportions may be obtained in separate samples (reflecting this variability).

Understanding sampling also requires students to identify three different types of distributions (Castro Sotos et al., 2007; Harradine et al., 2011):

1. 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; the parameter of interest is the population proportion *p* of elements sharing a given property.
2. The distribution of sample data collected from the population where we compute the proportion of successes, a summary statistic of the sample, which is used to estimate the population parameter *p*. Each student's quantitative response to a task in our research is a four-value sample from a binomial distribution.
3. The sampling distribution for the statistic, in this case the sample proportion. While the parameter *p* is an unknown constant, the sample proportion is a random variable that varies from sample to sample. As such, it is characterised by a probability distribution describing all possible values of it in samples of the same size that are selected from the population. In Section 3.2, the sampling distribution of the mean and the range for four-value samples in Task 4 are discussed.

Since there is a one-to-one correspondence between the *number* and the *proportion* of successes in a sample of *n* elements and since number is the easier concept, it is appropriate for students in this research to be asked to provide probable values for the number of successes in various samples. 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.

2.2. UNDERSTANDING SAMPLING REPRESENTATIVENESS AND VARIABILITY

The abundant research on intuitive understanding of sampling started within the heuristics-and-biases programme by Kahneman and Tversky (as summarised in Kahneman et al., 1982), where *heuristics* are conceived as unconscious actions guiding the resolution of complex tasks. Such heuristics simplify probability problems but often lead to a bias in reasoning. In our research, we ask the students for a probable value of the sample proportion in samples of different sizes. In such a task, the following heuristics may apply:

- The *representativeness heuristic* appears when a subject only considers the similarity between the sample and the population in making a probabilistic judgment (Tversky & Kahneman, 1974). 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 individual result. An associated bias is the insensitivity to sample size when judging the variability of the sample proportion.
- The *recency heuristic* describes a tendency in giving a priority to the past sample results over information about the population. In the gambler's fallacy (Kahneman & Tversky, 1972), the subject believes the result of a random experiment will affect the probability of future events. We speak of positive recency if the subject assumes the upcoming results will reproduce the observed pattern, and negative recency when the expectation is the future results will compensate for the observed results.
- The *availability heuristic* bases an estimation of the probability of an event solely on the facility to recall examples of similar situations (Kahneman et al.,1982; Reber, 2017).
- The *equiprobability bias* considers results of any random phenomenon as equally likely (Lecoutre, 1992). The inclination to use this argument is even higher if there are only two outcomes possible.

Batanero and Borovcnik (2016) summarise these and further strategies in probabilistic situations: availability, equiprobability bias, control of the future, representativeness, anchoring, patterns, and personal experience and information. The authors identify difficulties in taking into account these strategies in teaching as the strategies seem to be applied unconsciously. Research related to heuristics is also summarised in Batanero et al. (2016), Jones and Thornton (2005), Pratt and Kazak (2018), and Shaughnessy (2007). Shaughnessy 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 when taking five consecutive samples (with replacement) of ten candies. Using variations of this task with students of different ages and countries, *a progression* in students' reasoning has been identified (Noll & Shaughnessy, 2012; Shaughnessy et al., 2004):

1. *Idiosyncratic* students base their prediction of sample variability on irrelevant aspects of the task (e.g., preference, physical appearance).
2. *Additive-reasoning* students tend to predict the samples taking only absolute frequencies into account.
3. *Proportional-reasoning* students 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.
4. *Distributional-reasoning* students connect centres and variability when making the predictions of sampling variability. In addition to reproducing the proportion, their arguments also reflect random variation.

As in previous research cited in this section, the tasks used in our study require students to predict the outcome of multiple samples; the difference between this study and the aforementioned studies lies in the type of analysis, the use of binomial distributions, and the systematic variation of the parameters of the distribution to investigate their effect on students' understanding.

A different approach is taken by Findley and Lyford (2019) who interview eight college students while solving tasks that are designed to elicit their ideas about the shape of the sampling distribution, the likelihood of different values, or the judgement of the similarity of the sampling distribution with the population distribution (which are only similar if the population is normally distributed). Although this is a promising research line, the goals and methods are not related to the present research, which has primarily the goal to investigate whether the students achieve a balanced view between the mean (the tendency) and the variability of the sampling distribution. We use a much larger sample of students and a more in-depth method of analysis in order to classify their answers statistically. Our students are also younger. Finally, we try to identify strategies that influence the students' responses by analysing and structuring the arguments, they use to justify their answers.

2.3. UNDERSTANDING THE FREQUENTIST VIEW OF PROBABILITY

In the frequentist view, based on the Law of Large Numbers, the probability of an event is estimated from the relative frequency of this event in a long series of independent trials. Ideas of sampling representativeness and variability are implicit in this meaning where the precision of estimates depends on the sample size. Consequently, it is possible to interpret some research related to understanding the frequentist view of probability (e.g., Briand, 2005; Savard, 2010; Serrano, 1996) in terms of sampling. Below we summarise two investigations from which we adapted the tasks used in our research.

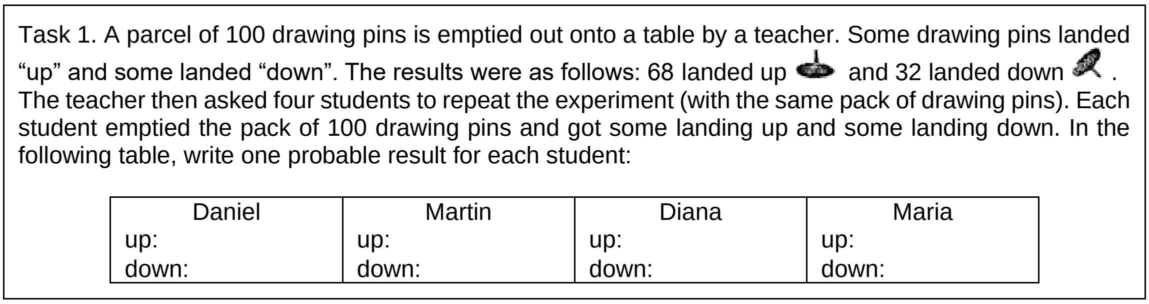
In his study on probability reasoning with children, Green (1983) set a problem where the children (11–16 years old) were given information about the result of a previous experiment and then had to predict the number of drawing pins landing *up* in emptying a parcel of 100 drawing pins (thumb tacks). Green found that 64% of the children reasoned according to the equiprobability bias (Lecoutre, 1992) since they asssumed that about 50% of the pins would land *up*. Similar results were found by Cañizares (1997) in a study with Spanish children (11–14 years old), only 17% of which provided a correct prediction.

The same task was used by Gómez et al. (2014) in a study with 202 prospective primary school teachers in Spain. They adapted the task to ask for predictions about four different trials in order to assess students' intuitions of representativeness and that of variability simultaneously. In our research, we use this new version of the task. Additionally, we vary the size of the sample and the probability of the event of interest systematically to investigate the impact of these variables. We also ask half of the students to justify their responses so we can apply a more complete analysis than in previous research. A preliminary version with a reduced sample (n = 127) and a simpler analysis was presented in Begué et al. (2018). While Begué and colleagues used a quantitative analysis of responses and the population distribution to classify the students' responses, in this paper we expand the sample size and use the sampling distribution of the mean and the range to classify the students' responses. We also add a qualitative analysis of justifications given by a subgroup of the students to better describe the students' reasoning.

3. METHOD

3.1. SAMPLE, TASKS, AND VARIABLES OF INTEREST

A total of 234 high-school students (17–18 years old) from five different schools, two in Huesca and three in Zaragoza (Spain), representing 13 class groups, took part in the study. The students had studied the curricular content described in the introduction during the six previous years of secondary school including sampling, sampling distribution, and computations with the sampling distribution. They were given a questionnaire including four tasks with the same format and questions, the first of which is reproduced in Figure 1 and the remaining in the Appendix. We also asked 127 of these students (half of the students, selected at random) to justify their responses (about half the students in each group and school). Each student entered the answers to the different tasks into a form provided by the researcher. In Task 1, taken from Gómez et al. (2014), 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.



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

Figure 1. Example of task given to the students

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). This distribution describes the number of times a given event happens in an experiment situation with n identical and independent trials, each of which with only two possible results (in Task 1, each simple experiment is observing the landing of one pin). The binomial distribution applies in this situation whether the experiments are carried out at the same time (emptying the full parcel of pins at a time) or successively (each pin is thrown one after another). Since p is unknown we expect students to use the proportion $\hat{p} = 0.68$ (which is an unbiased estimator of p with miminal variance; see Zacks, 2014), given in the stem of the question to estimate p. We used the data provided in the original task by Green (1983), who only asked for one probable result, in order to compare our results. Of course, the number of pins landing up depends on the material, the weight of the pin head, and the ratio of diameter to length of the pin. In our experience, performing the experiment with our students (different from those in the sample), the number of pins landing up were always higher than those landing down, and were similar to the numbers given in Task 1 for the type of pins we used. Since we asked for probable results, we expected that students would suggest values close to the distribution mean so we could evaluate their perception of that mean. At the same time, the four values provided by each student would assist us understand their perception of variability in sampling.

In Table 1, we include the parameters for all tasks. The context of the different tasks were: emptying drawing pins on a table (Task 1), flipping fair coins (Tasks 2 and 3), and throwing balls to a basket as is done in basketball (Task 4). The context of each task was deliberately varied from ideal random games to an experiment, which reflects properties from physics, and finally to a game where the skill of throwing the ball to the target introduces a human factor. In order to study the impact of sample size on students' answers, the sample size was varied from 10 to 100.

Table 1. Summary of task features

Task feature	Task 1 B(100, 0.68)	Task 2 B(100, 0.50)	Task 3 B(10, 0.50)	Task 4 B(10, 0.70)
Sample size	100	100	10	10
Population proportion	0.68	0.50	0.50	0.70
Expected number of successes	68	50	5	7
Standard deviation for the population	4.66	5.00	1.58	1.45

3.2. QUANTITATIVE ANALYSIS

Once the questionnaires were collected, we performed a statistical analysis of the four responses provided by each student for each task. The mean of the four values provided by each student was used to evaluate their intuitive understanding of the relationship between the population and sample proportions while the range of these four values served to assess their understanding of sampling variability.

Classification of the mean of the four predictions The number of successes, *X*, follows the binomial distribution B(*n*, *p*), which has an expected value of *np* and a standard deviation of $\sigma = \sqrt{np(1-p)}$. Therefore, we considered that the student had a good intuitive understanding of the mean for the sample proportion if the mean value was close to *np* (theoretical mean).

The mean was also compared to central intervals of the theoretical sampling distribution (whose standard deviation in the sample of four values is $\sigma/\sqrt{4} = \sigma/2$) with the interpretation of *optimum* understanding if the student's mean lies within ± $\sigma/2$ around the theoretical mean and acceptable if it is within ± $2\sigma/2$. Values not contained in the latter interval indicate a significant deviation (5%) from the usual understanding of the centre of the sampling distribution. For Tasks 1 and 4, where the binary outcomes in the experiment are not equiprobable, we additionally investigated students' answers falling into an interval of ± $\sigma/2$ around $p = 1/2$ to determine whether the students reason according to equiprobability, that is, they estimate the mean close to 50% of the sample size. From a normal approximation of the theoretical distribution of the mean of four binomial data, the intervals suggested should have a probability of approximately 68% and 95%. We used the normal approximation to describe an easily accessible classification of the data of the students. By simulation we could find a high agreement: the intervals derived in our approximate procedure coincide with the results of a computer simulation of the sampling distribution (we used 15,000 samples with four data points according to the parameters of the related task).

Classification of the range of the four predictions A simple rule based on the normal approximation to the binomial distribution for the distribution of the range in four predictions (see below) was derived from the data. Yet, a direct simulation yields more precise results than a general, but approximate, rule. Thus, we performed a simulation for the distribution of the range of four values of the number of successes in the binomial distribution B(*n*, *p*) for each task and found an empirical sampling distribution for the range, which is asymmetric. See Figure 2 for the result of the simulation of B(100, 0.68), which applies in Task 1. Again, for each task we computed the intervals containing 68% (optimum range) and 95% (acceptable range) in the empirical sampling distributions for the range. Notice that the simulation provides only an estimation of the theoretical distribution. We performed 15,000 simulations so that the percentile values we use to classify the students' answers are quite precise (they became stable only after 5,000 simulations).

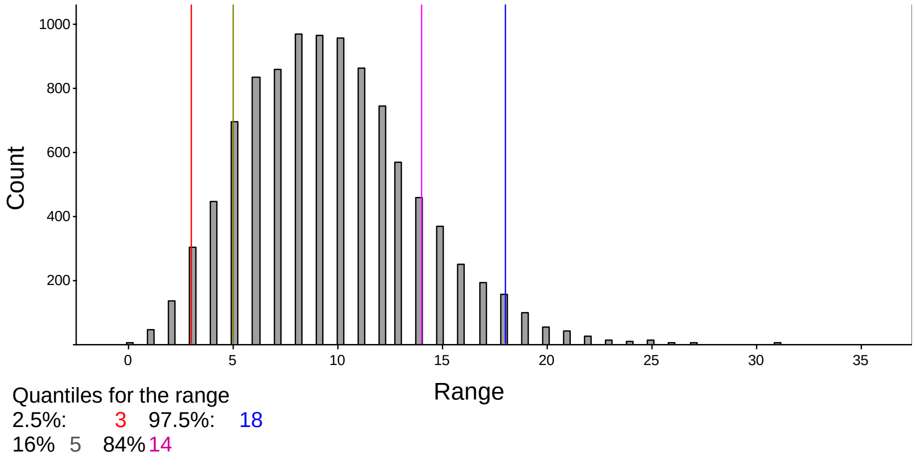


Figure 2. Simulated sampling distribution of the range for four-data samples of B(100, 0.68)

In summary, Tables 2 and 3 show the calculated intervals for each of the tasks for both the mean and the range of the four tasks (for the latter, we estimated the required quantiles from the simulation scenario). The classes based on these nested intervals are meant to be the interval in the pertaining row, but excluding the interval in the row above. We consider a response as optimum when the mean (or the range) of the four predictions provided by the student falls inside the 68% interval of the sampling distribution and acceptable when it falls inside the 95% interval.

Table 2. Classification for the interpretation of the mean by task

Classification *Mean of 4 data*	Interpretation *Mean is*	Task 1 B(100, 0.68)	Task 2 B(100, 0.50)	Task 3 B(10, 0.50)	Task 4 B(10, 0.70)
Lowest 2.5%	Too small	< 63.3	< 45.0	< 3.4	< 5.6
Centre 68%	Optimum	65.7–70.3	47.5–52.5	4.2–5.8	6.3–7.7
Centre 95%	Acceptable	63.3–72.7	45.0–55.0	3.4–6.6	5.6–8.4
Upper 2.5%	Excessively large	> 72.7	> 55.0	> 6.6	> 8.4

Table 3. Classification for the interpretation of the range by task

Classification *Range of 4 data*	Interpretation *Range is*	Task 1 B(100, 0.68)	Task 2 B(100, 0.50)	Task 3 B(10, 0.50)	Task 4 B(10, 0.70)
Lowest 2.5%	Too small	0–2	0–2	0	0
Centre 68%	Optimum	5–14	6–15	2–5	2–4
Centre 95%	Acceptable	3–18	3–20	1–6	1–6
Upper 2.5%	Excessively large	19–100	21–100	7–10	7–10

An approximate rule to classify ranges in samples of Binomial distributions The boundaries of the different intervals for the classification of the range in Table 3 was obtained by simulation: they provide a good intuitive picture of the classification of the quality of the perception of variability of a student. Rescaling these interval boundaries by the standard deviation of the Binomial distribution B(*n*, *p*), which is equal to $\sigma = \sqrt{np(1-p)}$, we find the following approximate rule for the range of the four values:

- The response is optimum if the range falls within the interval (1.2σ, 3σ); these ranges appear with a frequency of 68% in the sampling distribution.
- When the range of four values is within the interval (0.6σ, 4σ) but not in the previous interval, it is considered as high yet still acceptable; this interval contains 95% of the sampling distribution of the range.
- The range is judged as too small (students misperceive variability) when the range is below 0.6 σ; these ranges only occur in 2.5% of times.
- Ranges larger than 4σ are judged as excessively large (2.5%).

These criteria were used to classify the students' responses in other tasks where we vary the parameters *n* and *p*. Note that the classification for the range is independent of the population mean and that the percentage of cases in the intervals may vary a little because the variables are discrete.

3.3. QUALITATIVE ANALYSIS

To analyse the students' answers, we performed a content analysis (Krippendorf, 2013) of the detailed written arguments provided by the 127 students who were asked to justify their responses. Content analysis is a qualitative method where textual data is transformed into variables and categories to make inferences about the text content. We used one variable for the justification in each task (four variables in total) with a common set of categories. The categories were defined inductively from the data taking into account different types of reasoning in sampling described in previous research and summarised in Section 2. To establish the categories, we performed the following steps:

1. We recorded the responses of each student (who was assigned a code) to each task.
2. We compared all responses and grouped similar arguments in an initial set of categories.
3. This set of categories was progressively refined after several revisions of responses by the researchers and discussion of the coding with other colleagues.

4. RESULTS

4.1. UNDERSTANDING THE EXPECTED VALUE

We first investigate how well the students perceived the central tendency (in this case the probability of success) of the experiment. If students had a good understanding of central tendency, the mean of their predictions should lie close to the expected value. The summary statistics for the observed mean of the four predictions by each student and for the theoretical sampling distribution for the four tasks is presented in Table 4. To facilitate the comparison between students' answers and what is theoretically to be expected, we include central intervals containing 68% and 95% of these distributions. These two intervals indicate a frame where the sampling distribution for the mean lies; these nested intervals also separate the categories for the classification of the answers of the students (see Table 2). The students' answers should fit into this frame if they followed the theoretical expectations.

Table 4. Summary statistics for the theoretical sampling distribution and the observed mean of each students' four predictions

	Task 1 B(100, 0.68)		Task 2 B(100, 0.50)		Task 3 B(10, 0.50)		Task 4 B(10, 0.70)	
	Observed	Theor.	Observed	Theor.	Observed	Theor.	Observed	Theor.
Mean	57.6	68.0	51.2	50.0	5.1	5.0	6.6	7.0
95% interval	31.2–73.5	63.3–72.7	30.5–62.7	45.0–55.0	3.1–6.8	3.4–6.6	0.0–8.3	5.6–8.4
68% interval	48.7–69.0	65.7–70.3	46.5–54.5	47.5–52.5	4.5–5.5	4.2–5.8	5.0–7.0	6.3–7.7

General understanding of the expected value The distribution of the means of the student responses suggests (in general) a good understanding of the relationship between the population and sample proportions by the participants, given the proximity between the theoretical mean in the population and the mean value of the distribution of all students' responses in Tasks 2 to 4 (Table 4). In Task 1, however, there is a difference of more than 10 percentage points between the theoretical value[1] (68.0) and the mean of all values provided by the students ($\bar{x}$ = 57.6).

In Tasks 2 and 3, the mean of the four answers of the students is well coordinated with the theoretical distribution as the central 68% intervals of the students are completely embedded within the theoretical 68% intervals (see Table 4). The 95% interval of the observed data, however, is wider than the theoretical interval in Task 2. The situation is different for the experiments with non-equally-likely results (Tasks 1 and 4): the central (68%) interval is shifted downwards; in Task 4, it spreads a little beyond the limits and this also happens with the 95% intervals of the observed data from the students. This shows a systematic deviation in comparison to the theoretical sampling distribution.

In Figure 3, we display the distribution of the mean of the four predictions by each student in the different tasks. In Tasks 2 and 3 where the events in the experiment are equally likely, the responses are concentrated around the theoretical value of 50%; that is, 50% of students' answers lie in the centre of the distribution (the central box located between the quartiles). This does not happen in the experiments with non-equally-likely events, where the lower quartiles fall outside the 68% interval in Task 4 and even outside the 95% interval in Task 1.

[1] The theoretical value p = 0.68 of Task 1 expressed as percentage, coincides as number with the probability of values that should be included the central intervals signified by dashed lines in Figure 3 (68%). These two numbers should not be confused.

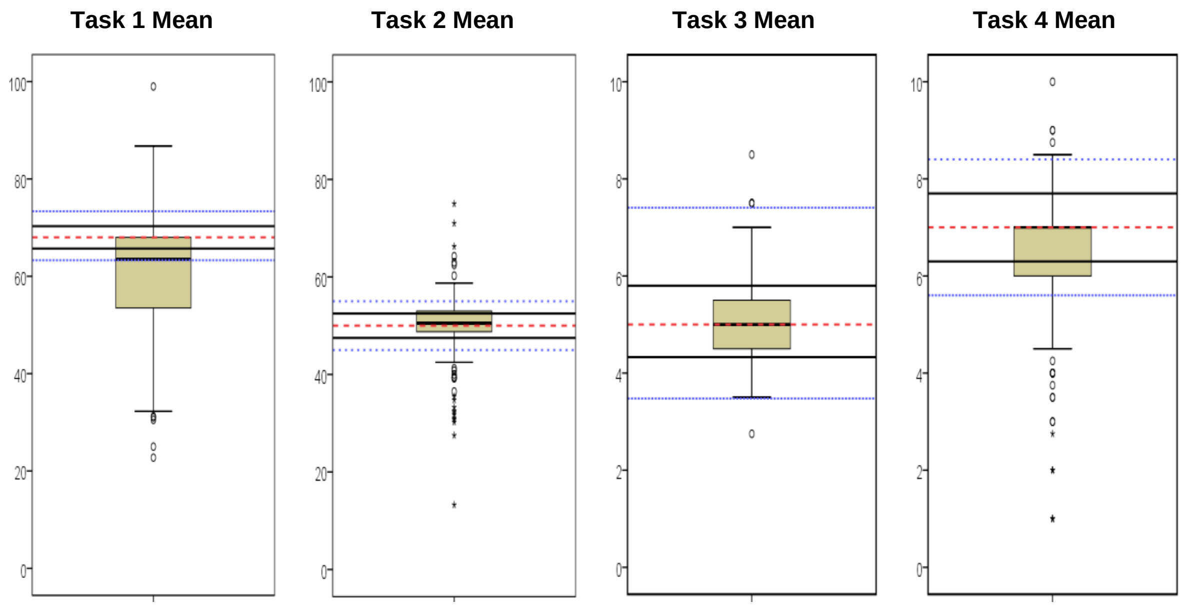


Figure 3. Distribution of the mean of each students' four estimates: dashed red line: theoretical means; solid black line: theoretical intervals containing 68%; dotted blue line: 95% of the sampling distribution for the mean of four predictions

Reasoning behind the general understanding The information from Table 4 is expanded in Table 5, where we reproduce the percentage of students providing different types of response in each task. The percentages in Table 5 are meant to sum to 100% as the labels in the first column represent non-overlapping categories of students' answers. Thus, we have to eliminate from the 95% interval the data in the 68% interval to form the category of "Acceptable". Likewise, we remove the data falling in the category "Equiprobable" from the interval for the category of "Below acceptable".

Table 5. Percentage of students with a mean of their predictions falling into different intervals of the theoretical sampling distribution

	Task 1 B(100, 0.68)		Task 2 B(100, 0.50)		Task 3 B(10, 0.50)		Task 4 B(10, 0.70)	
Response	Interval	%	Interval	%	Interval	%	Interval	%
Optimum (68%)	65.7–70.3	21.9	47.5–52.5	50.2	4.2–5.8	75.1	6.3–7.7	62.4
Acceptable[1] (95%)	63.3–72.7	24.5	45.0–55.0	22.3	3.4–6.6	13.7	5.6–8.4	6.6
Below acceptable	< 63.3[2]	32.1	< 45.0	15.9	< 3.4	8.2	< 5.6[2]	18.6
Above acceptable	> 72.7[3]	3.0	> 55.0	11.6	> 6.6	3.0	> 8.4[3]	2.2
Equiprobability bias	45.0–55.0	18.5	–	–	–	–	4.5–5.5	10.2
All		100.0		100.0		100.0		100.0

[1] Outside the 68% interval.
[2] These students reason by negative recency in our interpretation.
[3] These students reason by positive recency.

From Table 5, it follows that most students provided values whose mean falls either in the interval that theoretically contains 68% of values for the sampling proportion or in the interval that contains 95% of the values (optimum and acceptable responses) in Tasks 2 to 4. The number of students with optimum *and* adequate responses in Task 1 is high (46.4%) yet not as good as in the other tasks (72.5%, 88.8%, 69.0%). It is worthy to note that the higher percentages of optimum responses appear in Tasks 3 and 4 that correspond to small samples of the binomial distribution.

We also observe students with a specific kind of bias. One group of students showed a negative recency bias and another group a positive recency bias in Task 4 with $p = 0.68$ (negative 18.6%, positive 2.2%) and in Task 1 with $p = 0.70$ (negative 32.1%, positive 3.0%). Another group did not consider the frequency information about the event of interest and provided predictions with a mean very close to 50% (10.2% in Task 4 and 18.5% in Task 1).

4.2. UNDERSTANDING SAMPLING VARIABILITY

We now investigate how well the students perceive the variability of the experiment by analysing the range of the four predictions the students made in each task. The summary statistics for the observed range of the four predictions by each student and for the theoretical sampling distribution in the different tasks are presented in Table 6. A range that is too small indicates that the student focuses too narrowly on the tendency, while a range too large indicates that the student overrates the inherent variability. Again, we compared the central intervals containing 68% and 95% of the sampling distribution of the range of the four predictions by each student with the theoretical sampling distribution of ranges (Table 6). These intervals separate the categories for the classification as optimum, acceptable, below or above acceptable (see Table 7).

Table 6. Observed and theoretical summary statistics for the range of each students' four predictions

	Task 1 B(100, 0.68)		Task 2 B(100, 0.50)		Task 3 B(10, 0.50)		Task 4 B(10, 0.70)	
	Observed	Theor.	Observed	Theor.	Observed	Theor.	Observed	Theor.
Mean	$\bar{r}$ = 23.61	$\bar{R}$ = 9.58	$\bar{r}$ = 22.05	$\bar{R}$ = 10.25	$\bar{r}$ = 3.49	$\bar{R}$ = 3.22	$\bar{r}$ = 2.86	$\bar{R}$ = 2.94
95% interval	0–82	3–18	0–70	3–20	0–8	1–6	0–8	1–6
68% interval	7–40	5–14	5–45	6–15	2–5	2–5	2–5	2–4

General understanding of variability We start with a comparison of the mean of the sampling distribution of the range and the observed mean for the distribution of ranges of the four values provided by each student, which are presented in Table 6. The data alludes to difficulties in the perception of the variability in sampling. While for large samples, students' mean range is more than twice the mean value of the sampling distribution (23 and 22 as compared to roughly 10), for small samples these values nearly coincide (both values are a little above or below 3). The extremes of the 68% and 95% intervals in the theoretical sampling distribution of the range[2] are much wider than the 68% and 95% intervals of the observed ranges for the four predictions provided by the students in Tasks 1 and 2, and are closer to each other in Tasks 3 and 4.

In Figure 4, the distribution of the range of the four estimates by each student in the different tasks is displayed to which we added the mean (dashed line in red) and intervals with the central 68% of ranges (solid line in black) and the central 95% of the simulated distribution of ranges (dotted line in blue). These distributions suggest that the perception of the variability in sampling is not equally good for $n = 100$ and $n = 10$ as we observe a high percentage of students whose ranges are located above the upper point of the 95% interval, both in Tasks 1 and 2 corresponding to big samples. On the contrary, the centre of the distribution is located within the 68% intervals in Task 3 and 4 corresponding to small samples.

[2] These intervals were estimated by a simulation study (see Section 2).

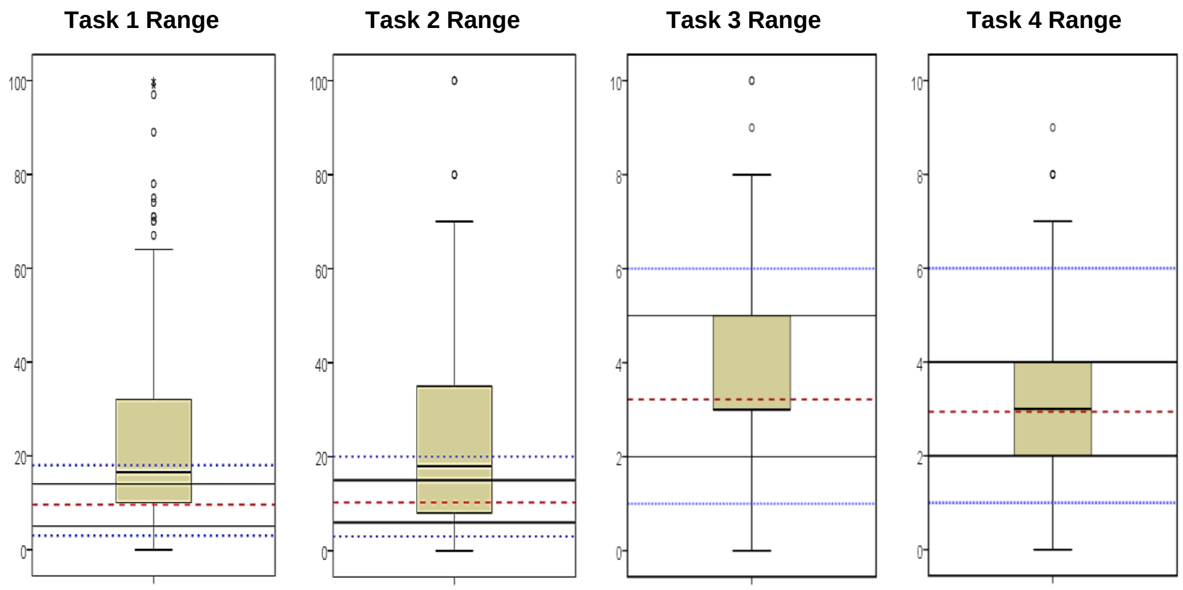


Figure 4. Distribution of the range of each students' four estimates: dashed red line: theoretical means; solid black line: theoretical intervals containing 68%; dotted blue line: 95% of the sampling distribution for the mean of four predictions

Reasoning behind the general understanding This information is completed in Table 7, where we reproduce the percentage of students giving different types of response. The optimum answer has a percentage of less than a quarter (20.5%) for Task 1 and a third (31.6%) for Task 2. For Tasks 3 and 4, the optimum range was attained by a much larger proportion of the students (72.6% and 51.7%). Taking the optimum and acceptable class together, the percentages are 42.7, 47.0, 77.7, and 67.5 (from Task 1 to 4). As regards the extreme categories: the lowest 2.5% of the sampling distribution accounts for 12 to 16% of students with no marked distinction between small and large samples while the upper 2.5% accounts for 44.9 and 38.9% of the answers for n = 100 but only roughly 7% for n = 10. That means, nearly one half (Task 1) and more than one third (Task 2) of the students' ranges are classified as upper extreme. All these data reflect a better overall perception of the inherent variability in small samples as compared to large samples.

Table 7. Percentage of students with ranges classified in different intervals

Classification of the range	Task 1 B(100, 0.68)		Task 2 B(100, 0.50)		Task 3 B(10, 0.50)		Task 4 B(10, 0.70)	
Response	Interval	%	Interval	%	Interval	%	Interval	%
Optimum (68%)	5–14	20.5	6–15	31.6	2–5	72.6	2–4	51.7
Acceptable [1] (95%)	3–18	22.2	3–20	15.4	1–6	5.1	1–6	15.8
Lowest 2.5%	0–2	12.4	0–2	14.1	0	14.5	0	15.8
Upper 2.5%	19–100	44.9	21–100	38.9	7–10	7.7	7–10	7.3
All		100.0		100.0		100.0		100.0

[1] Outside the 68% interval

4.3. STABILITY OF THE RESPONSES TO THE DIFFERENT TASKS

In order to study whether students' predictions in the different tasks are related, we computed the correlations between the mean value of each student's four predictions in the different tasks (Table 8), and the correlations between the range of each student's four predictions for each task (Table 9).

We obtained positive correlations of middle or high size for all tasks in both statistical summaries. As regards the means (Table 8), this result suggests a tendency of the same student to produce predictions of related values in all tasks; that is, that most students produce high, middle size, or small predictions (as compared to the expected value in the task) in *all* the tasks, across the context and the sample size. In particular the highest correlation appears between Tasks 1 and 2, which indicates a *similar type of response to both tasks* by many students.

Table 8. Pearson's correlation between the mean of the four predictions in each task

	Task 1 Mean	Task 2 Mean	Task 3 Mean	Task4 Mean
Task 1 Mean	1	0.612**	0.418**	0.356**
Task 2 Mean		1	0.481**	0.371**
Task 3 Mean			1	0.474**

** p< 0.001

Table 9. Pearson's correlation between the range of the four predictions in each task

	Task 1 Range	Task 2 Range	Task 3 Range	Task 4 Range
Task 1 Mean	1	0.708**	0.444**	0.469**
Task 2 Mean		1	0.531**	0.435**
Task 3 Mean			1	0.493**

** p< 0.001

The correlations between the ranges of each student's four predictions for the different tasks (Table 9) are still more intense and again positively correlated for all tasks. This means that when a student provided four predictions with a small range in one task, the student tended to repeat values with small ranges in all the other tasks; the same "co-ordinated" behaviour occurs for medium or high range. Again, the highest correlation appears between Tasks 1 and 2, which suggests a tendency to provide the *same type of variability in big samples*. Since the percentage of students giving responses with high variability (Table 7) in these tasks was about 40%, these correlations confirm our interpretation that an essential proportion of students in our sample tended to overestimate the variability in big samples.

As a result of the correlation analysis, we can state that the behaviour of the students was very stable; a stability that seems even stronger within the same sample size presented in the tasks, at least for the tendency, a little less for the range, which captured students' perceptions of variability.

4.4. ANALYSIS OF STUDENT'S INDIVIDUAL ARGUMENTS

In this section, we investigate the arguments provided by the randomly selected 127 students who were asked to justify their responses in order to understand their reasoning better and corroborate the previous findings from the in-depth statistical analysis of the four predictions of each student in each task. We analysed students' answers with the aim of finding strategies they use implicitly by structuring the arguments used to justify their answers. We present categories, which emerged after several rounds of refinement in the content analysis of the answers. The categories are described theoretically and then are illustrated by examplary responses of our study. A synthesis of the results concludes this section.

Randomness The student refers to the chance character inherent to the situation and suggests that the result cannot be predicted because of randomness. An example follows where student S442 suggested that there cannot be a pattern because of chance and provides values that are not too close to the centre of the distribution.

> S442 in Task 3: All depends on chance; there is no pattern so that I provide random results (6, 4, 3, 8).

Physical properties of the device but no probabilistic language The student describes physical features of the random generator to justify the response yet does not mention probability. For example, student S473 took into account the weight of the pin and imagined the way it would move until it lands *up* but used no probability language. Thus, he predicted more pins landing *up* in three samples although, inconsistenly, in the last sample he decided to use 50%. Two of his values were above and two were below the mean so that he seemed to compensate results in different trials. This type of reasoning involves an underlying heuristic of representativeness (Tversky & Kahneman, 1982) and a lack of understanding of the independence of the trials. Interestingly, his predictions decreased systematically and the student used predictions only in steps of five.

> S473 in Task 1: I considered the weight of the pin base; when you throw it, it will fall *down* but it will rotate and equilibrate so that the pin tends to land *up*. Then, in my response I suggested that more pins will land *up* (75, 70, 60, 50).

Physical properties of the device and probabilistic language The student considers the physical characteristics of the generator and uses probabilistic language at the same time as in the following example. We interpret that the student's probabilistic reasoning is more complete as he mentions probability explicitly. For example, Student S527 provided three predictions that lie all above the given data. In forming his response, he used arguments related to physics and probability much more than the data (frequencies) given. Again, only steps of five occur in the predictions.

> S527 in Task 1: I selected these results since there is much more probability that the pin lands *up* because of its shape and the head's weight (70, 60, 75, 80).

Frequentist probability The student's response is based on the experimental data provided in the task formulation. This response is interpreted as an intuitive understanding of the frequentist approach to probability, according to which the data in the new sample should be close to that probability. In the following example, the student establishes a proportion of success using the data in the task and, consequently, he predicts results close to this value although with some variability. Student S412 uses his everyday knowledge of the context as a good player should have an almost constant achievement.

> S412 in Task 4: The player succeeded in 70 out of 100 trials. If he throws 10 balls to the basket, then the result will be 7 successes. In case he tries 40 times, the result should be 28 (7, 6, 8, 7).

Classical probability The student mentions the possible and the favourable cases in the experiment and assigns a probability using Laplace's rule assuming that all cases are equally likely. Sometimes this procedure involves an equiprobability bias (Lecoutre, 1992) and other times a correct proportional reasoning of students (Shaughnessy et al, 2004). In the following example, the student used this as well as the argument of variability described below. Student S457 seemed to be very close to equity but showed too much confidence in his predictions; hence, he underestimated the inherent variability.

> S457 in Task 2: We assume that the results are equally likely so that each result has 50% probability. Nevertheless, an exact result will be unlikely so that I have predicted some results close but different to 50 (48, 53, 51, 50).

Convergence and variability The student uses arguments based on the close connection between the population proportion and the sample proportion as well as on the variability of the sample proportion due to a random sampling process (Batanero & Borovcnik, 2016). This reasoning suggests a distributional reasoning in the students according to Shaughnessy et al. (2014). The next example shows these arguments, which is combined with physical features of the device and probabilistic language (Argument 3). Student S434 alternateed the predictions above and below the data given and there seems to be an equiprobability bias involved as the deviations from the data are much smaller in the upper than in the lower direction.

> S434 in Task 1: The teacher got 68% of pins landing *up* and 32% landing *down* so that the percentages obtained by the children should be similar. In moving a bit these percentages, e.g., by 8%, we obtain 60-74%, and so on but we always need to remember the weight of the pins so that it is more likely that the pin will land *up* (70, 60, 72, 65).

Explicit equiprobability bias The student explicitly mentions equiprobability of the events due to the randomness in the experiment but this equiprobability is not appropriate. Student S470 follows the equiprobability bias (Lecoutre, 1992), predicting results very unlikely for the binomial distribution; in his response, he included values in the whole range of the distribution. Apart from the first value that is close to the data given, this student seemed to have an awkward perception of equiprobability as he switched between the extremes of 0 and 100.

> S470 in Task 1: Any possible result is correct because there is a 50% probability that the pin will land *up* and a 50% probability that the pin will land *down*. Hence, the teacher's results are wrong; the results have nothing to do with those produced by the teacher (73, 2, 100, 0).

Subjective beliefs The student supports his or her response by other criteria not related to probability as in the following example where the student re-interprets the situation in a deterministic way. Assumptions are made that the different trials are not related to each other as they are not different trials of the same experiment but are related to different experiments due the multiplicity of factors that may affect the result. This reasoning reflects a perception of the task not related to probability. The situation may also be perceived as random but with a much smaller variation than in the binomial experiment as the skill of the player should establish a "smooth" performance.

> A470 in Task 4: Is different to the other tasks. Here, there is technique in addition to chance so that the result could be good or bad depending on the day. If the player is comfortable with the game, his motivation, etc. (5, 8, 8, 7).

Discussion of the categories of arguments used The frequency of arguments in the different tasks is summarised in Table 10 and Figure 5. Some students were classified into more than one category as they essentially used more than one argument. We observe an association of some arguments with different tasks where the students perceived the task features and reasoned mainly according to them. In Task 1, the most frequent argument is the physical feature of the device (either with or without probabilistic language) that occurs in more than 40% of the students' answers. This argument is justified in this task as there is a tendency in the pin to land *up*. The fact that students often used this argument supports the conclusion that the students predominantly applied a propensity view of probability in this task; that is, they conceived probability as a measure of a physical disposition of a random system (the pin here) to behave in a certain way. This argument is missing in Task 4, which has a much higher prevalence of subjective beliefs (about the experiment or by perceiving the situation as a non-random experiment) than the other tasks (17%). Tasks 2 and 3 are highly loaded on classical probability (more than 40%) but attract also many convergence arguments (more than 30%) and still many randomness arguments (more than 20%).

Table 10. Percentage of students providing different arguments in each task

Type of Argument		Task 1	Task 2	Task 3	Task 4
		B(100, 0.68)	B(100, 0.50)	B(10, 0.50)	B(10, 0.70)
1.	Randomness	12.6	21.3	22.8	6.3
2.	Physical properties of device, no probabilistic language	9.4	1.6	0.0	0.8
3.	Physical properties of device and probabilistic language	33.1	8.7	8.7	0.0
4.	Frequentist probability	27.6	5.5	3.1	56.7
5.	Classical probability	2.4	40.2	43.3	1.6
6.	Convergence and variability	7.9	32.3	31.5	40.2
7.	Explicit equiprobability bias	7.9	5.5	3.9	0.8
8.	Subjective beliefs	7.1	4.7	3.1	17.3
9.	Gives no justification	14.2	27.6	26.0	22.8

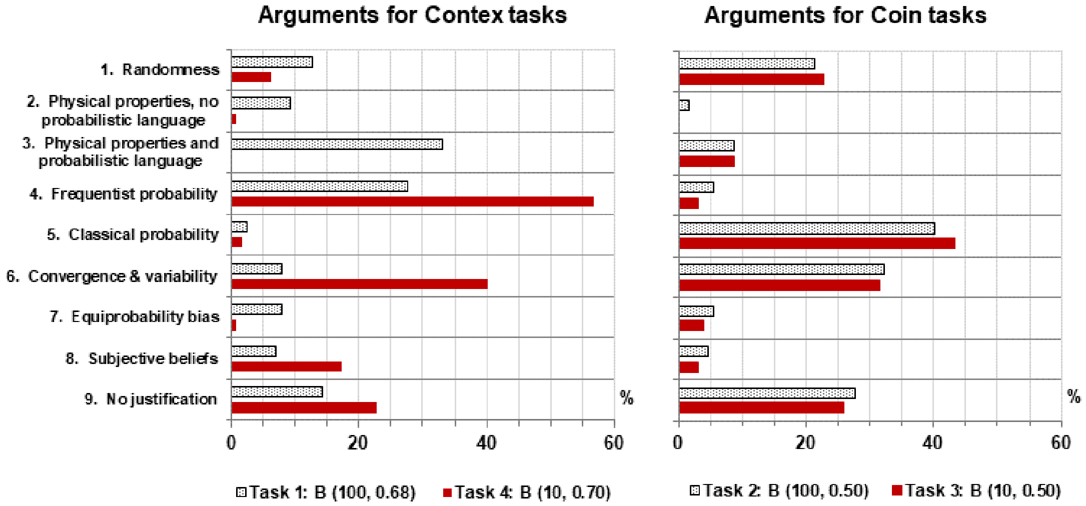


Figure 5. Percentage of students providing different arguments in each task

An intuitive application of frequentist probability is more common in Tasks 1 and 4 where the outcomes of the experiment have different probabilities, while in Tasks 2 and 3, corresponding to equiprobable events, students tended to reason according to classical probability. If we add students using a propensity meaning of probability (Arguments 2 and 3), the majority of students used their implicit – partially correct – ideas of probability to justify their responses.

The arguments also corroborate the existence of biased reasoning in the students. Although there are few students who explicitly show the equiprobability bias (roughly 1 to 8% depending on the task), some of the responses classified as Argument 1 (randomness, or anything may happen because it is random) may induce this bias implicitly and increase the prevalence of the equiprobability bias. Other students based their justifications on subjective beliefs (3–7%, yet 17% in Task 4) and a substantial percentage of students did not provide any justification (14–27%, varying with the task). Arguments that reflect students' concerns solely about sampling variability are linked to the idea of convergence (8% in Task 1 and 30–40% in the other tasks). In these arguments, students suggested that the frequency of results should be similar to that in the data of the task but, due to the random features of sampling, some variation should be expected.

This type of reasoning takes into account both expectation and variability; it corresponds to *distributional reasoning of sampling*, which was named and described by Shaughnessy et al. (2004). Distributional reasoning was very scarce in Task 1 where students tended to reason according to the physical features of the device and did not pay much attention to issues of convergence or variability; convergence, however, was well-covered in the other tasks (between 30–40%). In summarising, distributional reasoning was well covered in our study, the students aligned their predictions according to the tendency *and* the variability of the results simultaneously.

5. DISCUSSION

In our research, we used four variations of a task from Gómez et al. (2014), who adapted the task from Green (1983). Our study pursued a different purpose; instead of focusing on the perception of frequentist probability, as previous authors did, we centred our investigations on the sampling distribution for the proportion. We added three further tasks in order to investigate the influence of the parameters in the binomial distribution on the students' responses by changing the values of the paramaters systematically. Below we discuss the main results.

5.1. PERCEPTION OF THE EXPECTED VALUE

The mean of all predictions provided by the students was very close to the theoretical value in all tasks, except in Task 1 ($\bar{x}$ = 57.6), where the observed value in our study is very close to the value of 57.7 obtained by Gómez et al.(2014). The explanation for this result is that a proportion of the students considered both landing positions of the drawing pins as equiprobable so that they provided predictions for the number of pins landing *up* close to 50%. In Tasks 2 and 3 with equally likely events (the ideal coin), the responses are concentrated around the theoretical value of the sampling distribution.

We observed different biases in the students' reasoning about the mean. Some of the students used negative recency, in trying to compensate the data of a past experiment by a much lower success rate in the future. To our interpretation this would be compatible with their perception of variability. Other students used positive recency and then increased the success rate in their predictions (this can also be explained as their expression of variability). Finally, a third group did not consider the frequency information about the event of interest; their reasoning may be interpreted as equiprobability bias, according to Lecoutre (1992). In our study, it was not possible to attribute this behaviour to the composition fallacy, which is described by Chernoff and Russell (2012) (i.e., to transfer a property of the population to the sample) in Tasks 1 and 4 because in these populations the events are not equally likely.

5.2. PERCEPTION OF VARIABILITY

The differences between the mean of the sampling distribution of the range and the observed mean for the distribution of ranges of the four predictions provided by each student in the different tasks suggest that the students assumed too much variability in large samples as compared to small samples. This interpretation is confirmed by the percentages of students providing different values of ranges in the upper 2.5% intervals in Tasks 1 and 2 as compared to Tasks 3 and 4.

These findings contradict results from previous research. On the one hand, students do not disregard the sample size as it is assumed by the representativeness heuristics (Tversky & Kahneman, 1974), in the sense that they grossly overrate the variability of big samples but they do recognise the variability of small samples properly (see averages of the observed ranges of students' answers and the theoretical mean of the sampling distribution of the range in Table 6) and, consequently, they implicitly take into account the sample size in judging sampling variability. Yet, there is a non-negligible difference visible in the average range of four-sample answers between students' answers and theoretical values from the sampling distribution. The difference of our results as related to research dealing with the representativeness heuristics is rooted in the different task that is used. While in that research, people are asked to compare the variability of a small and a big sample from the same binomial *distribution*, in our study we ask the students to provide four different *samples* in each task and we draw conclusions on their reasoning from the statistical analysis of the data in these samples. Our conjecture is that students are more familiar with small samples so that they perceive the variability in experiments better with small samples than they do in large samples.

From our results, for big samples it is not clear that students reach the distributional reasoning level of sampling as described by Shaughnessy et al. (2004). According to these authors, at that level of reasoning, students consider mean and variability of samples in conjunction; in our study, the number of students who simultaneously consider mean and variability is much higher in small than in big samples. Apparently then, the rule for classifying a student to that level of reasoning depends on the sample size. We also remark that, compared to what is accepted by statistical theory, students consider the variability to be relatively higher in big than in small samples, which replicates findings by Gómez et al. (2014).

5.3. ARGUMENTS USED TO JUSTIFY THE RESPONSE

Some students link randomness to unpredictability, a conception described by many researchers, such as Briand (2005), Fischbein et al. (1991), Savard (2010), and Serrano (1996) in their studies of students' conceptions of randomness and probability. Serrano suggests that in this mode of reasoning, a student interprets a probabilistic question (and provides a probable result) in a non-probabilistic way (providing a result that will happen with certainty).

A second justification was based on physical properties of the device with or without using probabilistic language. There is an underlying conception of probability close to that of propensity introduced by Popper (1959) as a measure of the tendency of a random system to behave in a certain way and as a physical disposition to generate an outcome of a certain kind. This view of probability was also proposed by Peirce (1932/1910), to whom a probability generator (e.g., a coin) bears a tendency for its various possible outcomes that are both directly related to the long run and indirectly related to singular events. Batanero et al. (2016) resume Popper's term of propensity and use propensity in the long run as tendency to generate relative frequencies with particular values and conceive propensity in a single experiment as identical to the probability for an event.

In Tasks 1 and 4, other students predicted results close to the result of the first experiment given in the task, although with some variability; this phenomenon is also reported by Green (1983), Cañizares (1997), and Gómez et al. (2014). Other students applied classical probability in Tasks 2 and 3, which is a correct reasoning.

Finally, we also found students who justified the equiprobability of the events by the random character of the experiment even though this was not appropriate in the task, which suggests an incidence of the equiprobability bias (Lecoutre, 1992). Others expressed subjective beliefs or a deterministic view of the experiment described in Task 4, a type of response, which is described in Cañizares (1997), Fischbein et al. (1991), and Serrano (1996).

6. FINAL REMARKS

One difficulty in empirical studies designed to investigate individuals' intuitions about sampling (e.g., Green, 1983; Cañizares, 1997) is the subjects may conflate perceptions of mean (tendency) and variability (Chance et al., 2004; Rubin et al., 1991; Saldanha & Thompson, 2002; Watson & Moritz, 2000). If we ask them to provide a set of expected results, they tend to ignore the random character of the experiment and provide values too close to the mean (as happened with some students in Shaughnessy, 2007). In fact, some students in our study repeated the same value in all four predictions. On the other hand, when students rely too much on the random character of the situation, they may assume that the expected value would never occur (or at least only with a low, negligible probability) so they provide results with high variability, as occurred in parts of our study.

The use of different values for the parameters in the situation where the students had to provide *four* predictions, and the particular analysis applied in our research served to separate the aspects of tendency and variability and represent a novel approach with new results that add to existing research. One remarkable unexpected result of the present study is that the perception of variability seems to be much better for small samples than for larger samples. Therefore, the design of the experiment asking four predictions rather than one prediction or one judgement helped us to assess students' answers related to their perception of the mean and the variability simultaneously with respect to the sampling distribution of proportions. The innovative feature of the analysis in this study is to classify optimum and acceptable responses of the four values provided by the students, by taking into account the theoretical sampling distribution of the *mean* and the *range* (of four predictions) while previous studies only considered the binomial distribution *in the populatio*n.

A final contribution of the present study is the qualitative analysis of the justification of the responses in part of the sample, which suggests students differentiate the task features and apply many correct ideas about sampling, including the classical, frequentist, and propensity view of probability. Other students, however, reasoned according to biases such as the equiprobability bias (Lecoutre, 1986), or subjective beliefs described by Cañizares (1997), Fischbein et al. (1991) and Serrano (1996). Consequently, not all students reached the distributional level of reasoning about sampling described by Shaughnessy et al. (2004). Of course, as in any research, our tasks are limited and different tasks (e.g., changing the context, the number of trials, or the wording of the tasks) might lead to some variation in the findings. Consequently, all these results need further clarification and therefore we plan to continue this type of research in the near future with other samples, new tasks, and potentially, different approaches to the analysis.

Formal teaching of sampling and binomial distribution in Spain, based on solving textbook problems and learning definitions of concepts, did not help the students in our study to overcome their inadequate intuitions and reasoning. A consequence is the need to supplement teaching with tasks similar to those analysed in this paper. A classroom discussion of students' responses and arguments can be also supported by simulations of the experiments described in the tasks in order to make the students aware of the binomial distribution and key properties of samples of it.

ACKNOWLEDGEMENT

Project PID2019-105601GB-I00 (AEI) and research groups FQM126 (Junta de Andalucía) and S119 (Gobierno de Aragón, FEDER).

REFERENCES

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

Batanero, C., Chernoff, E., Engel, J., Lee, H., & Sánchez, E. (2016). *Research on teaching and learning probability.* ICME-13. Topical Survey series. Springer.
[Online: https://doi.org/10.1007/978-3-319-31625-3_1]

Borovcnik, M. (2012). Multiple perspectives on the concept of conditional probability. *Avances de Investigación en Educación Matemática, 2*, 5–27.
[Online: https://doi.org/10.35763/aiem.v1i2.32]

Briand, J. (2005). Une expérience statistique et une première approche des lois du hasard au lycée par une confrontation avec une machine simple [A statistical experience and a first approach to chance laws in high school by confronting students with a simple tool]. *Recherches en Didactique des Mathématiques, 25*(2), 247–281.

Begué, N., Gea, M. M., Batanero, C., & Beltrán, P. (2018). Do high school students understand the sampling distribution of a proportion? In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics*. International Statistical Institute.
[Online: https://iase-web.org/Conference_Proceedings.php?p=ICOTS_10_2018]

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). Springer.

Cañizares, M. J. (1997). Influencia del razonamiento proporcional y combinatorio y de creencias subjetivas en las intuiciones probabilísticas primarias [Influence of proportional and combinatorial reasoning and subjective beliefs in primary probabilistic intuitions]. Unpublished doctoral dissertation. Universidad de Granada.
[Online: www.ugr.es/~batanero/pages/ARTICULOS/CANIZARE.pdf]

Castro Sotos, A. E., Vanhoof, S., Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98–113.

Chance, B., del Mas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. B. Garfield (Eds), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Springer.

Chernoff, E. J., & Russell, G. L. (2012). The fallacy of composition: Prospective mathematics teachers' use of logical fallacies. *Canadian Journal of Science, Mathematics and Technology Education, 12*(3), 259–271.
[Online: https://doi.org/10.1080/14926156.2012.704128]

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). Springer.

Findley, K., & Lyford, A. (2019). Investigating students' reasoning about sampling distribution through a resource perspective. *Statistics Education Research Journal, 18*(1), 27–46.
[Online: https://iase-web.org/documents/SERJ/SERJ18(1)_Findley.pdf?1558844313]

Fischbein, E., Sainati Nello, M., & Sciolis Marino, M. (1991). Factors affecting probabilistic judgements in children and adolescents. *Educational Studies in Mathematics, 22*(6), 523–549.

Gómez, E., Batanero, C., & Contreras, J. M. (2014). Conocimiento matemático de futuros profesores para la enseñanza de la probabilidad desde el enfoque frecuencial [Prospective teachers mathematical knowledge to teach probability from a frequentist approach]. *Bolema, 28*(48), 209–229.

Green, D. R. (1983). From thumbtacks to inference. *School Science and Mathematics*, *83*(7), 541–551.

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). Springer.

Heitele, D. (1975). An epistemological view on fundamental stochastic ideas. *Educational Studies in Mathematics*, *6*(2), 187–205.
[Online: https://doi.org/10.1007/BF00302543]

Huerta, M. P. (2018). Preparing teachers for teaching probability through problem solving. In C. Batanero & E. Chernoff (Eds.). *Teaching and learning stochastics: Advances in probability education research* (pp. 293–311). Springer.

Jones, G. A., & Thornton, C. A. (2005). An overview of research into the teaching and learning of probability. In G. A. Jones (Ed.), *Exploring probability in school* (pp. 65–92). Springer.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–453.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. 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]. *Boletín Oficial del Estado, 3*(1), 169–546.

Noll, J., & Shaughnessy, J. M. (2012). Aspects of students' reasoning about variation in empirical sampling distributions. *Journal for Research in Mathematics Education, 43*(5), 509-556.

Peirce, C. S. (1932). Notes on the doctrine of chances. In C. S. Peirce, *Collected papers* (Vol. 2, pp. 404-414). Harvard University Press (Original work published in 1910).

Popper, K. R. (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science, 10*(37), 25–42.

Pratt, D., & Kazak, S. (2018). Research on uncertainty. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 193–227). Springer.

Reber, R. (2017). Availability. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory* (p. 185–203). Routledge.
[Online: https://doi.org/10.4324/9781315696935]

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* (pp. 314–319). International Statistical Institute.
[Online: 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*(3), 257–270.

Serrano, L. (1996). *Significados institucionales y personales de objetos matemáticos ligados a la aproximación frecuencial de la enseñanza de la probabilidad* [Institutional and personal meanings of mathematical objects linked to the frequentist approach to teaching probability]. Unpublished doctoral dissertation. Universidad de Granada.
[Online: www.ugr.es/~batanero/pages/ARTICULOS/TESISSERRANO2.pdf]

Savard, A. (2010). Simulating the risk without gambling: Can student conceptions generate critical thinking about probability? In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics*. International Statistical Institute.
[Online: https://iase-web.org/Conference_Proceedings.php?p=ICOTS_10_2018]

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957-1009). Information Age Publishing.

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 Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 177–184). PME Group.
[Online: www.emis.de/proceedings/PME28/RR/RR045_Shaughnessy.pdf]

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

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. 84–100). Cambridge University Press.

Watson, J. M., & Moritz, J. B. (2000). Development of understanding of sampling for statistical literacy. *The Journal of Mathematical Behavior, 19*(1), 109–136.

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

CARMEN BATANERO
University of Granada (Spain)
batanero@ugr.es

APPENDIX: TASKS 2 TO 4 GIVEN TO THE STUDENTS

Task 2. A parcel of 100 fair coins is emptied onto a table by a teacher with the following result: 53 Heads and 47 Tails. The teacher then asked four students to repeat the experiment with the same parcel of coins. Each student emptied the pack of 100 fair coins and got some Heads and some Tails. In the following table, write one probable result for each student.

Elena	Clara	Matías	Rosa
Heads:	Heads:	Heads:	Heads:
Tails:	Tails:	Tails:	Tails:

Task 3. A teacher asks 4 students to flip 10 coins on the table and count the number of Heads and Tails. In the following table, write one probable result for each student.

Silvia	Javier	Miguel	Carmen
Heads:	Heads:	Heads:	Heads:
Tails:	Tails:	Tails:	Tails:

Task 4. From 100 attempts of a basket-ball player to throw the ball into the basket from the free-throw line, 70 are shots (land in the basket). In the following table, write one probable result for four games in which he throws the ball 10 times from the free-throw line.

Game 1 (10 throws)	Game 2 (10 throws)	Game 3 (10 throws)	Game 4 (10 throws)
Number of shots:	Number of shots:	Number of shots:	Number of shots:
Number of failures:	Number of failures:	Number of failures:	Number of failures: