Repeated sampling in a digital environment: A remix of data and chance

Marianne van Dijke-Droogers[1], Paul Drijvers[2] and Arthur Bakker[3]

Utrecht University, The Netherlands

[1]m.j.s.vandijke-droogers@uu.nl; [2]p.drijvers@uu.nl; [3]a.bakker4@uu.nl;

Drawing statistical inferences (SI) is essential in a society where data are of increasing importance. Understanding the relation between data and chance, necessary to understand statistical inference, is however challenging for students. Technological innovations – such as the Sampler in TinkerPlots (TP) – enable students to investigate this relationship by modeling a population and simulating repeated samples. Along this line, the research reported here presents the results of a pilot with fourteen 9th-grade students, inexperienced with sampling, in a Learning Lab. The pilot focuses on how students use TP as a digital environment for exploring data and chance – i.e. what strengths and constraints they encounter – and how they subsequently use this information for SI. The results suggest that the participating students encountered difficulties in modeling the population, however, they were able to simulate, explore and reason inferentially through repeated sampling in TP.

Keywords: TinkerPlots, repeated sampling, (informal) statistical inference, modeling, statistics education

Drawing statistical inferences (SI) is essential in a society where data play an increasingly important role. However, the handling of variation and uncertainty involved in drawing inferences based on sample data, is challenging for students (Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007; Konold & Pollatsek, 2002). New technological innovations – such as the Sampler-option in TinkerPlots – enable students to model and investigate the interaction between data and chance in order to gain insight into variation and uncertainty (Biehler, Frischemeier, & Podworny, 2017; Pfannkuch, Ben-Zvi, & Budgett, 2018). Following this approach, Van Dijke-Droogers, Drijvers, and Bakker (2018) suggested an approach of repeated sampling with a black box filled with colored balls to introduce students, inexperienced with sampling, to the key statistical concepts of sample, frequency distribution from repeated samples and the simulated sampling distribution. This approach with a black box seemed promising to invite students to investigate the intertwined relationship between data and chance. As a follow-up to this study, the research reported here investigates *how* activities in the digital environment of TinkerPlots can strengthen the statistical insights of students, i.e. the interrelation between data and chance.

Theoretical background

Statistical inference

Statistical inference includes making statements about an unknown population based on observed sample results. In contrast to descriptive statistics, which concerns describing the data under investigation, inferential reasoning includes the handling of sampling variation and interpreting the role of chance.

Digital environments for statistical modeling

Recent digital environments that use dynamic visualizations and offer opportunities to create statistical models, enable students to investigate the intertwined relationship between data and chance. Tools such as TinkerPlots that have both data analysis and sampling capabilities, support young students to think about the modeling process. Creating statistical models to simulate data from repeated sampling can be helpful to develop key statistical ideas of distribution and sampling variation (Konold, Harradine, & Kazak, 2007). In an instant of time, a model of the population can be built in a digital environment and then be used to generate a large number of (repeated) sample results. In this way, student can explore sampling variation due to chance, especially when they visualize the results of repeated sampling in a sampling distribution. In this way, students engage in modeling, dealing with variation and thinking about the context (Pfannkuch et al., 2018). The work by Garfield, delMas and Zieffler (2012) shows that students can learn to think and reason statistically – or as the authors call it "really cook" – by using statistical modeling in digital environments.

Repeated sampling in the black box activity

This learning—or cooking—effect is reflected in the work by Van Dijke-Droogers et al. (2018) on repeated sampling from a black box filled with small colored balls ("balletjes" in Dutch). The results of that study suggest that repeated sampling from a black box is a promising approach to introduce students to the concepts of sample, frequency distribution on repeated sampling (resampling) and the simulated sampling distribution, and, therefore, invites reasoning about variation and uncertainty involved in SI. An important characteristic of this 3-step learning approach seems the strong connection between a physical black box experiment in step 1, the visualization of expected sample results from resampling with a black box in a frequency distribution in step 2, and the simulated sampling distribution on resampling in the digital environment of TinkerPlots in step 3. An overview of the learning steps in the black box activity is displayed in Table 1.

1. Repeated sampling with the physical black box
2. Visualization of expected sample results from repeated sampling with the black box in a frequency distribution, by manually drawing a sketch
3. Simulation of a large number of samples from repeated sampling with the black box and visualization of results in the sampling distribution, by using the Sampler-option in TinkerPlots
4. Using the digital environment of step 3 in a new context

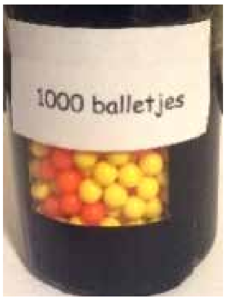


Table 1: Overview of learning steps in the black box activity

Van Dijke-Droogers et al. (2018) focused on step 2 and showed that students could, based on their experience in step 1, imagine and sketch a frequency distribution from repeated samples with most sample results close to the population proportion and in which strong deviations hardly occur. As a next step, the aim of the research presented here, is to gain insight into *how* statistical modeling with the Sampler-option in TinkerPlots can strengthen the statistical insights of students concerning sampling variation due to chance. Therefore, we added a fourth step to the black box activity to investigate *how* students use TinkerPlots to explore data and chance in new situations and *how* they use this information for SI.

Methods

To investigate how students explore and reason with TinkerPlots, we conducted a pilot study in the Teaching and Learning Lab of Utrecht University.

We have chosen to run the pilot in a lab setting. The advantage of this lab setting, over a classroom environment, was that detailed recordings could be made of the actions and conversations of several teams of students working on the same assignment at the same time. During a five-hour session, seven 9th-grade students worked in teams of 2 or 3, on tasks addressing steps 1 to 4 of the learning trajectory. This pilot was repeated in a similar five-hour session, with seven other students. The teams worked on a laptop, the screen of which was displayed on an interactive whiteboard. Camera recordings of their actions on the screen were made, and student conversations were recorded. Students noted their findings as a team on a student worksheet. We worked with students from pre-university level, the 15% best performing students in our educational system. Figure 1 shows the setup in the Learning Lab.

Figure 1: An impression of the setup in the Learning Lab

During the pilot, students went through learning steps 1–4 of the black box activity. Special focus was on how students applied their digital experiences from step 3 to investigate the role of variability and chance in the new context of step 4. In step 3, students used the sampler option in TinkerPlots to simulate repeated samples from the context of the black box. While doing so, they used an instruction sheet with the necessary technical actions in TinkerPlots. In this step, students investigated possible sample results by modeling a black box filled with 750 yellow and 250 orange balls at different sample sizes and different number of repetitions. During a group session in step 3, students discussed the boundaries of common sample results and decided that the middle 80% of the results can be regarded as “most common sample results”. They indicated the 10% highest sample results as “exceptionally high” and the 10% lowest sample results as “exceptionally low”. It was expected that the students would apply their experiences from step 3 in the new context of step 4, as detailed in Table 2. This table shows an overview of components A to G with statistical insights, TinkerPlots activities in step 3 and expected students’ activities in step 4.

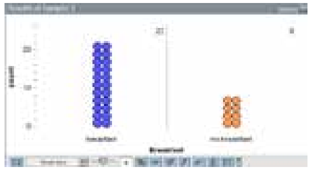
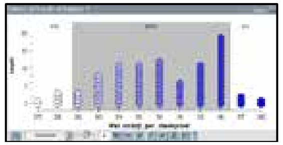
Components of statistical insights		TinkerPlots activity in step 3 with the black box	Expected students' activity in TinkerPlots at step 4
A	Sample results vary due to chance	Context: A black box is filled with 750 yellow and 250 orange balls. For each sample, the number of yellow balls is counted. Examine the sample results you can expect.	Context: At the beginning of the school year 210 out of the 300 students on a specific primary school were used to having breakfast daily.. Task 1: To investigate whether this is still the situation at the end of the school year, a sample of 30 students will be taken. Examine which sample results you can expect if the breakfast habits of students are unchanged.
B	A model of the population can be entered and examined in a digital environment	Model the population in TinkerPlots by using the Sampler-option. The model can be entered as a pie chart, bar graph, histogram, dot plot or curve.	Students enter a model of the population in the same way as described in the cell on the left (TP activity B)
C	Sample size effects sampling variation due to chance	Enter the sample size you want to examine.	Students enter the sample size they want to examine. In this example, the sample size is 30.
D	Sample data can be listed in a table A visualization of sample data in a plot provides a clear overview	Simulate one sample, the results are automatically displayed in a table. To make a plot of the sample data, explore the data and choose a visualization that provides a clear overview.	Students simulate one sample of which the data are automatically listed in a table. They use the Plot-options to explore and clearly visualize their data. 
E	Repeated sampling in a digital environment can produce a large number of sample results. These results can be listed in a table by choosing one specific characteristic.	Use the history button to have repeated samples memorized. To do this, choose a specific characteristic—for example the number of yellow balls—that you want to record from each sample.	Students use the history button to record the number of "students who had breakfast" from each sample.
F	A visualization of multiple sample results in a plot can be used to estimate possible sample results	Visualize the sample results in a sampling distribution. Use dividers to examine possible sample results, i.e. most common sample results, exceptionally high or low	Students vizualise the sampling distribution and use dividers to examine possible sample results 
G	A larger sample size reduces the variation in the corresponding estimates of the population and hence, leads to a better outcome.	Estimate most common sample results and extrapolate these to the population. Then, try different sample sizes later for comparison. Draw a conclusion about the effect of larger sample sizes to the estimate of the population	Students estimate most common sample results and exceptionally high or low results to draw their inferences corresponding to the task (other sample sizes were addressed in tasks 2–4)

Table 2: An overview of statistical insights, TinkerPlots activities in step 3 and expected students' activities in step 4

Sample size 30	**Expected number of students having breakfast daily** (interval notation)
Most common sample results	
Exceptionally high results	
Exceptionally low results	

Table 3: Table on student worksheet

Task 2: In the past year a lot of attention has been paid to a 'healthy' breakfast at the school. At which sample result, sample size 30, is it likely that the breakfast habits of students have improved? Support your answer with data.

Task 3: At the end of the school year, a sample of 30 students was not feasible. Therefore, a sample will be taken of only 10 students. With which sample result can you now assume that the breakfast habits of students have probably improved?

Task 4: On closer inspection, the school board decides to postpone the sample to the next school year and to draw a sample of 100 students at that time. For which sample result is it likely that the breakfast habits of students have improved?

Task 5: Compare the sample results and corresponding estimates of the population at different sample sizes. What can you conclude about the population estimates from a larger sample size?

Table 4 Tasks 2-5 on Student Worksheet

For Task 1 in step 4, students could use a table on their worksheet, as displayed in Table 3. Students used TinkerPlots to determine most commons sample results. For Tasks 2 to 5, more in-depth questions that mainly focused on the role of sample size were asked within the same context, as shown in Table 4. Students were free to choose their own working method within TinkerPlots.

Results

The analysis of the video-recordings shows that the expected students' activities in components A and C to G did occur. However, the students encountered difficulties with component B. Students started with discussing the breakfast context in component A. Students seemed to be involved in this real-life context as they exchanged and discussed all kinds of possibilities and expectations based on their personal experiences and intuition, for example "Only 70% of these students have breakfast, that's way too low for a healthy school. Here, we have a lot more students who have breakfast." After sharing their thoughts, they started working in TinkerPlots. Here, in component B, making a model of the population in TinkerPlots, the difficulties began as 3 out of 6 teams used the sample information for their model. As such, they modeled the population by entering 21 for breakfast and 9 for no breakfast. When simulating samples from this population they got confused and asked the teacher for help. Referring to the black box activity in step 3, made them aware of the difference between population and sample. Another difficulty with the modeling was that 2 teams used a pie chart with percentages. These students entered 70% for breakfast and 30% for no breakfast. Although they used similar proportion as the given population, working with percentages refers to an endless population and in this case the population size is given with 300. So, a finite population of 300 is a better model. After struggling at component B, the other components were performed as expected, although students occasionally needed guidance from the teacher, which mainly consisted of linking the new context in step 4 to the known context of the black box activity in step 3.

The activities in TinkerPlots (re)strengthened the statistical insights of students. As a first strength, each new component offered opportunities for students to discuss and reason about their actions in the digital environment and the implication of their actions. For example, at component A, a student stated “Well, they should take a larger sample to get a better picture of the school” and “Probably 21 students of the sample will have breakfast, or at least about 21 students. 22 or 23 is possible too. Even 30 is possible, but I don’t think that will happen”. Statistical concepts such as the effect of sample size and possible sample results that were addressed in steps 1–3 of the black box activity, were again considered and discussed by students in a new context.

As a second strength, students referred to terms from steps 1–3 of the black box activity, as they were using terms like “yellow and orange balls” and “content” of the black box. The lay-out and set-up of the digital environment offered opportunities to easily relate both contexts. The following fragment shows an example of students’ reasoning while modeling the population in task 1 of learning step 4, where terms related to repeated sampling with the black box are underlined:

Student 1: Okay, we have a sample of 30. If nothing has changed, then probably 21 students will have breakfast.

Student 2: I think so…

Student 1: But it could be 22 or 23 students. I think that 27 or more cannot happen. What do you think?

Student 2: Of course, it is possible, but....

Student 1: What should we enter here (*pointing at the empty startup window of the Sampler*)? This is actually the total content, right? So, all students of the school?

Student 2: Why not the 30 students… we are supposed to check on 30 students

Student 1: I think, we should take the 300, because that are all students, like all the balls in the box and the 30 is just a sample, the visible ones… don’t you think?

Student 2: Okay, so then we must enter 210 with breakfast and 90 no breakfast.

A third strength concerned the easy and rich way in which students could explore the data. They could easily create and compare different graphs, like pie charts, dot plots, etc. On the other hand, these explorations took a lot of time and it seemed like students based their decisions, for example about the format of the graph, on personal ideas and less on clarity. Moreover, choosing the right colors and shapes for the balls, was time consuming.

A fourth strength was that every component required action in TinkerPlots. Students could not thoughtlessly enter their values, but they had to be alert and aware of their actions.

In step 4, students were asked to carry out five investigative tasks and to note their answers on student worksheet. An outline of the results on students' worksheet are displayed in Table 5.

Task	Results	Frequency in teams of students	Example of students' reasoning
1.	Correctly filled in Table	6 out of 6	-
2 .	Correct conclusion	6 out of 6	More than 25 students is exceptionally high, so …… probably the breakfast habits of students is improved. 24 students is the boundary of most common sample results. So, with 26 students or more, it can be assumed that the breakfast habits have improved.
3.	Correctly filled in Table	6 out of 6	-
	Correct conclusion	6 out of 6	With a sample result of 9 or 10 students (out of 10) that have a daily breakfast, we can assume that the breakfast habits are improved, because this is an exceptionally high number.
4.	Correctly filled in Table	6 out of 6	-
	Correct conclusion	5 out of 6 *(one team empty)*	When more than 80 students (out of 100) are having a daily breakfast, this is extremely high, and therefore it is assumable that more students are having a daily breakfast.
5.	Correct conclusion		A larger sample size gives a more precise estimate of the population The larger the sample size, the less variation in the corresponding estimate and the bigger the chance of a good estimate

Table 5: Overview of results on students' worksheet

Conclusion and Discussion

The research reported here focused on how activities in the digital environment of TinkerPlots can strengthen students' statistical insights, i.e. the interrelation between data and chance. The main difficulty of students was creating a correct model of the population and, more particularly, distinguishing sample and population characteristics from a given context. However, entering the characteristics into the tool, TinkerPlots, did not cause any problems. The awareness of the difference between sample and population can be improved by spending more time and instruction to modeling a population in a variety of contexts.

A first strength of the activity in the digital environment of TinkerPlots was that students exchanged and discussed statistical information which may lead to a deeper notion of the concepts addressed. A second strength was that the lay-out in the digital environment could easily be related to the black box activity (balls, sample size). Students explanations contained black box terms several times. A third strength concerned the convenient and rich way in which students explored the data. However, choosing appropriate representations was also time consuming and mainly based on personal preference over clarity. A fourth strength was that every thinking component required action in TinkerPlots, which increased their awareness of the statistical concepts involved.

Although the results seemed promising, a few critical points must be considered. The fact that the students went through the learning steps quite easily may have been caused by the short time frame in which the activities logically followed each other. Although the students were given quite open tasks in step 4, their approach was strongly influenced by their previous activities in step 1 – 3. In addition, the students occasionally needed help from the teacher, which consisted of linking the new context to the black box activity in step 1 – 3.

The results show that this activity with repeated sampling in the digital environment of TinkerPlots is a promising way of how the interaction between data and chance can be taught in a brief session. Researcher who would like to repeat this activity should consider that this research highlights the results of a small-scale pilot in a laboratory setting with high performing students.

References

Biehler, R., Frischemeier, & D., Podworny, S. (2017). Editorial: Reasoning about models and modeling in the context of informal statistical inference. *Statistics Education Research Journal*, *16*(2), 8–12.

Castro Sotos, A. E., Vanhoof, S., Noortgate, W. van den, Onghena, P. (2007). Students’ misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 1*(2), 90–112.

Garfield, J., Ben-Zvi, D., Le, L., & Zieffler, A. (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics, 88*(3), 327–342.

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education, 2*(1). Article 1. Retrieved from http://escholarship.org/uc/item/38p7c94r

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259–289.

Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance and context. *ZDM Mathematics Education, 50* (online).

Van Dijke-Droogers, M.J.S. van, Drijvers, P.H.M., & Bakker, A. (2018). Repeated Sampling as a step towards Informal Statistical Inference. 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.