COMBINATORIAL THINKING AS KEY FOR INTRODUCING HYPOTHESIS TESTING—EVALUATION OF THE PLANNED SECONDARY-SCHOOL REFORM IN HUNGARY

Péter Fejes Tóth[1], Ödön Vancsó[2], and Manfred Borovcnik[3]
[1] Hungarian University of Agriculture and Life Sciences, Department Applied Statistics, Budapest
[2] Eötvös-Lóránd University, Centre for Didactics of Mathematics, Budapest
[3] University of Klagenfurt, Institute of Statistics, Klagenfurt
fejes.toth.peter@uni-mate.hu

We developed an educational programme to prepare a curricular reform in Hungary. In the framework of this project, we tested and evaluated a new curriculum design relying on combinatorial and probabilistic knowledge and skills. As opposed to inferential statistics, which is not part of secondary education, combinatorial thinking has a rich tradition in secondary education. The idea was to extend combinatorially introduced probability to a wider meaning of general frequencies by using probability to evaluate the credibility of hypotheses. The materials proved to be interesting, motivating, and understandable. Based on pre- and post-tests, students' probabilistic knowledge improved significantly. It is worth noting that our approach also intensified their understanding about probability apart from the targeted content of hypothesis testing.

INTRODUCTION

Statistics education in Hungarian secondary schools is limited to descriptive statistics. However, there are efforts under way by a research group from the Hungarian Academy of Sciences to change the status quo. Internationally, a rich body of research literature is available regarding the methods, practice, and experiences of including inferential statistics in secondary curricula (Borovcnik, 2017; Garfield et al., 2008, 2012; Rossman & Chance, 1999). One way to expand the curriculum is by introducing statistical hypothesis testing, as has happened in numerous cases (starting from, e.g., Bunt, 1967). We built curriculum materials by relying heavily on the fact that, traditionally, there is a well-established combinatorial and probability tradition within the curriculum in Hungary. We conducted an experiment with one class (Fejes Tóth, 2020) and used the preliminary results to refine and compile a six-lesson curriculum. The curriculum was delivered by trained teachers in an experimental setting in four secondary school classes (grades 10–12) during the 2019–2020 and 2020–2021 school years (with some delays due to the pandemic). In this paper, we present the results of the experimental teaching phase, from the perspective of how already acquired literacy in probability can support the understanding of hypothesis testing and conversely, how mastering this statistical tool affects the understanding of probability. We introduce the intended curriculum, discuss teachers' and students' evaluations of the curriculum (regarding its usefulness and suitability), and provide an evaluation of its educational impact in terms of knowledge and problem-solving skills, with a focus on the effects on teaching probability observed in the experimental classes.

COMPILED CURRICULUM

In designing the curriculum, we heavily built on the approach of Tamás Varga (Varga, 1970, 1972, 1983) by focussing on research-based learning and learning-by-doing. We included several experiments and "hands on activities," and we tried to encourage students to perform calculations and draw conclusions as independently as possible. During the course, students started solving problems using prior knowledge with combinatorial calculations, and they were gradually introduced to applying their calculations to informally evaluate statistical hypotheses, which served to prepare them for hypothesis tests. We assessed hypotheses in terms of goodness of fit using the chi-squared test. We also pointed out that hypotheses may differ in nature and require different methods, including non-statistical methods, to test them. We chose three experiments as tasks for lessons that should enable students to understand the essence of probability theory and to shape an adequate understanding of hypothesis testing based on probability. The first two tasks can be solved with the probability knowledge students previously acquired. The formulation of the research question draws students' attention to hypothesis testing. The three experimental problems are discussed in more detail below. In all three experiments, students are guided through the entire process from formulating a hypothesis, conducting the

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

experiment, analysing the data, and drawing conclusions about the hypothesis. Type I and Type II errors are also introduced, emphasising the importance of choosing the level of significance.

The Lady-Tasting-Tea Experiment

Students work in pairs. One student fills four glasses with mineral water and four glasses with tap water. The other must determine which glasses are filled with tap water, knowing that out of the eight glasses, four contain mineral water and four contain tap water. The first student notes the number of correctly classified glasses. The two students then swap roles. Questions asked include: How can we decide whether the subject is actually able to identify the different types of water? Alternatively, someone recognised all four glasses of tap water. Can we say that the person recognises tap water? Overall, students have to decide, based on their experiment (tasting glasses of tap water or mineral water), whether they think the test person is able to distinguish between the two types based on taste.

Is the Coin Fair or Loaded?

Several tasks were given to students to decide whether a coin is regular or loaded for a specific event such as the outcome of 63 Heads and 37 Tails in 100 tosses of the coin, or the outcome of 58 Heads and 42 Tails. Students investigated a variety of cases using a prepared Excel spreadsheet with results for many tosses, few tosses, fake coins, etc. Although the calculations and simulations referred to the probability of some event, the task led students in the direction of informal tests of a statistical hypothesis. Students could not know whether the coin they had in their hands was regular or loaded. The simulations in Excel also provide an opportunity for students to investigate their rates of wrong decisions, either rejecting or not rejecting the hypothesis of a regular coin. During the experiment, students not only analyse their own experimental results, but are also given different distributions as examples to see how the assessment changes depending on sample size and distribution.

Is the Die Fair or Loaded?

Several tasks were given to students to decide whether a die is fair or loaded.

- Task 1. Each student uses a spreadsheet to simulate 60 rolls of a fair die. Students draw a bar chart based on the frequencies and make a subjective decision about whether the die is fair or loaded at a glance. They then calculate the chi-squared value to make a decision.
- Task 2. Each student simulates 60 rolls of a slightly loaded die. Again, the students first make a decision by looking at the graph and then applying the chi-squared test. Anyone who concludes that the die is fair made a Type-II error.
- Task 3. Repeat the experiment with more rolls (60, 600, and 6000 rolls). Students should notice that the Type-II error gets smaller with increased numbers of rolls. Then students repeat the experiment with a heavily loaded die and notice that the frequency of the Type-II error again decreases.
- Task 4. Students are each given a loaded or regular die, but that fact is unknown to them. Their task is to find out whether their die is loaded or not based on measurements and calculations.

For each task, students have to decide, based on their experiment (rolling a die several times, running simulations for series of 60, 600, and 6000 rolls) if the die is fair or loaded. In both cases (simulating or rolling a die), they work with fair *and* loaded dice (e.g., physical dice created with a 3D printer). They record their results numerically and in graphs and make decisions based on the visual appearance of the graphs and the application of chi-squared tests. The parameters that influence the probability of a Type-II error are also discussed.

The Collection of Tasks

The tea and coin problems can be solved using combinatorial probability calculations and using knowledge about sampling with and without replacement. The die problems can be solved by a best-fit test (chi-squared test). The coin problem can also be solved using a chi-squared test and is a good task to use to present a new method in addition to applying the binomial distribution that is already well known to students. Due to limitations both in mathematical knowledge and time, we cannot present the mathematical background of the chi-squared test to students, however, we can show that by using combinatorial probability calculations or using the chi-squared test, we usually come to the same

conclusion. For all three tasks, the students can perform physical experiments and amend them later using computer simulations, where it is also possible to base calculations on a larger sample size.

The tea and coin experiments are essential for illustrating our approach, which is based on the idea of building on students' prior knowledge of combinatorics by making a close connection to the new topic of statistical testing (they are complete novices with statistical inference) and by reinterpreting simple probabilities calculated in a familiar combinatorial setting. The tea experiment is special because the inference question arises directly from the context. It is a close replica of the original experiment run by biologist Ronald Fisher (Fisher, 1935), who used eight cups of tea to find out (test) whether the "Lady" can recognise whether milk was added to tea or the other way around. The Lady knew there were four cups of both kinds. We have to use the hypergeometric distribution for calculating the probabilities of interest. The probability that the Lady identifies all four tea-first cups correctly if she merely guesses equals $p = 1/8C4 \approx 0.0143$, i.e., the value of *p* cannot be lower than this value even if she identifies all four mugs correctly. If the Lady does not know the number of cups with milk added to the tea, we have to use the binomial distribution, and this value decreases to $p = 1/2^8 \approx 0.0039$. Students would need to consider different options: eight cups or more, known or unknown number of cups with "tea first," all guessed correctly or not. In different experimental settings, the possible cases are different, which would lead to a different evaluation of the evidence from the actual experiment to judge the hypothesis that the Lady has a special skill of recognising how the tea was prepared (see Fanshawe, 2021). The probability of classifying all cups correctly may vary from 0.00024 to 0.0143 depending on whether eight or twelve cups are involved and whether the Lady is told the number of each preparation. Based on prior knowledge, students can easily calculate the solution probabilities by applying the binomial and hypergeometric distributions known from probability and can develop understandings of the essence of hypothesis testing from the way the questions are formulated.

EVALUATION METHODS AND DATA

The pilot study serves to answer our main research question: *Are the suggested methods and problems appropriate to introduce the concepts of inferential statistics to high-school students for the very first time?* The aim of the pilot course was to test the feasibility, acceptability, and effectiveness of the course design among students and teachers. Four teachers took part and delivered the material to 64 students in their classes. We assessed students' and teachers' attitudes and perceptions regarding the design of the material and the lessons and assessed students' understanding, i.e., the learning outcomes (see Vancsó et al., 2018). We designed the assessment scheme for the study using mixed methods and repeated data collection. Summative classroom assessment (Suurtamm et al., 2016) was applied in a pre-test–post-test design to assess the impact of the programme in terms of skills and competencies acquired. The pre-test was administered before the pilot course and consisted of basic combinatorial questions. The post-test, administered after six pilot lessons, consisted of similar basic combinatorial questions (see Table 1) and a task requiring the application of freshly acquired knowledge about statistical hypothesis testing.

To assess feasibility and acceptability, students were asked to complete an anonymous questionnaire about their opinions and attitudes related to the material. The questions related to how well they understood the material, how interesting they found the lessons, and how useful they found what they had learned. Semi-structured interviews were conducted with the teachers and analysed in a narrative manner. The interview guide contained the questions listed in Table 2. In this paper, we present a segment of the evaluation phase that relates to probability education.

RESULTS

Pre- and Post-Test

For the pre-test, we had several assessment goals. First, we wanted to measure students' ability to decide whether a particular event in a specific setting is certain, possible, or impossible to happen. This task proved to be straightforward for the students: they had to make a decision regarding seven events, and the answers of the 59 students who completed the pre-test were 98.8% correct. Because we wanted to keep the pre- and the post-test as similar as possible, we asked similar questions on the post-test and got similarly good results. Second, we also measured students' combinatorial knowledge. Because we repeatedly used sampling with and without replacement in the new materials, we decided

to measure how proficiently students can handle such problems before and after the experiments. The difference here was striking. For sampling without replacement, 35.6% of students solved the problem correctly on the pre-test, whereas 83.6% gave correct answers on the post-test. For sampling with replacement, the success rate increased from 10.3% to 60.0%. The rate of completely wrong answers went down from values of approximately 60% to values around 15% for both types of tasks. The results between pre- and post-test differ significantly, as shown in Table 3.

Table 1. Questions in pre- and post-test

Type of task	Pre-test	Post-test
Sampling without replacement	*In an urn, there are 4 white and 11 non-white (5 red and 6 blue) balls. Take 5 balls without putting them back.*	*There are 100 apples in a box, 15 of which have worms. We select 5 apples randomly without replacement.*
	What is the probability that exactly 5 of the 5 balls are white? What is the probability that exactly 3 of the 5 balls are white?	What is the probability that there is no apple with a worm among the 5 selected apples? What is the probability that there are two apples with worms among the 5 selected apples?
Sampling with replacement	*Take out 5 balls with replacement (i.e., before you draw again, put the previously drawn ball back into the urn).*	*There are 100 apples in a box, 15 of which have worms. We select 5 apples randomly with replacement.*
	How many different results can we get if the order still does not matter? What is the probability that exactly 5 of the 5 balls are white? What is the probability that exactly 3 of the 5 balls are white?	What is the probability that there is no apple with a worm among the 5 selected apples? What is the probability that there are two apples with worms among the 5 selected apples?

Table 2. Interview questions for teachers

What inspired you to take part in the experiment? Have you received enough support from the teacher training? Was the used material or the teaching of it new to you?
What did you like best or least about the curriculum and the experimental teaching? Would you like to teach this material again in the future? If so, what would you change about it? If not, why not?
What do you think you can benefit from the practice you have gained? Has this experimental teaching influenced your relationship with the students?
In your opinion, did the students enjoy the lessons? Have they succeeded in learning the new material?
Is there anything you might wish to say, highlight, or share your experiences?

Table 3. Results on pre and post-test for combinatorial problems

	Sampling with replacement				*Sampling without replacement*			
	Pre-test		Post-test		Pre-test		Post-test	
	N	%	N	%	N	%	N	%
Completely wrong answer	36	62.1	9	16.3	38	64.4	7	12.7
Partly correct answer	16	27.6	13	23.7	0	0	2	3.7
Correct answer	6	10.3	33	60.0	21	35.6	46	83.6
Sum	58	100.0	55	100.0	59	100.0	55	100.0
	$\chi^2 = 35.148$, p-value $= 2.3\times10^{-8}$				$\chi^2 = 31.82$, p-value $= 1.7\times10^{-8}$			

Note: For sampling without replacement, the categories of *Partly correct* and *Correct* have been merged to provide a better approximation of the test statistic by a chi-squared distribution

On the pre-test, 17 of 59 students did not insert the binomial coefficient into the formula $p(x=k)=nCk\ p^k(1-p)^{100-k}$. On the post-test, this number reduced to seven. Though this was not a central part of the lessons, practice proved to have an impact in improving students' proficiency—this formula occurs frequently both in the "tea/water" experiment and in the "coin" experiment. The repeated application of a previously learnt concept in a "technical" setting enhanced learning. Theoretical understanding and practical application (ability to determine necessary steps to achieve certain results) create mathematical knowledge that is strongly connected by interrelations, possibly in an iterative manner (Rittle-Johnson & Schneider, 2015). This also means that the time allotted for the statistics course is not a detriment to other topics but should be regarded as an opportunity to deepen students' knowledge—different topics are not necessarily competing but can strengthen each other.

Beliefs and Evaluation of Teachers

The four teachers who participated in the experiment were fundamentally positive about the curriculum design and its content. One teacher noted, "the curriculum was interesting, especially because it was based on experiments carried out by students—in secondary schools this kind of hands-on activity is a rare phenomenon." A second teacher indicated, "even though hypothesis testing will not be part of the curriculum, I will definitely use the Lady-tasting tea experiment in my lessons on hypergeometric and binomial distributions."

These teachers recognised the curriculum as feasible both for them as teachers and for their students. The decision-making mechanism of hypothesis testing was well understood by everyone, as corroborated by students' post-test results. During the course of teaching the lessons, the teachers' experience was that students who had a better knowledge base in mathematics were more likely to have hard times digesting the concepts of a chi-squared test (used in the loaded-die experiment).

> It was more difficult for higher-performing students to accept the chi-squared-test method, which did not include a mathematical explanation. Actually, that bothered me too—as a teacher, I'm not used to not having answers. I understand that the underlying theory cannot be explained in a high school, yet it would be nice to provide a better, logical understanding.

The sophisticated mathematical details behind chi-squared tests were not explained in detail, and students with minds wired more for abstract mathematical concepts and understanding might have lacked the mathematics behind why the test works the way it works. The first problem (Lady-tasting-tea) was less of a problem for these students with stronger mathematical backgrounds (binomial and hypergeometric distribution), leaving them with no questions left open. In the case of students with a more moderate background in mathematics, the exact opposite was true. In the case of the chi-squared test, they were able to accept the method, mechanically complete calculations, and come to decisions without needing the mathematical background. While solving the tasks, these students found greater difficulty with using their (weaker) combinatorial knowledge and skills. In general, the teachers felt that they would like to use similar in-class experiments that can be solved based on combinatorial probability in the future, regardless of whether statistical hypothesis testing becomes part of the curriculum or not. They perceived these tasks to be predominantly of a probabilistic nature, with the question to answer or the answer itself stated in an unusual form.

CONCLUSION

Regarding the curriculum materials, it is obvious that preliminary knowledge in combinatorial probability and familiarity with combinatorial thinking makes the introduction of statistical hypothesis testing feasible. However, the six teaching units (45 minutes) were not enough for the curriculum to be realised, and we will address this in the next phase of our experimental class design. The results demonstrate that it is didactically helpful to base the introduction of a complex method for making decisions under uncertainty on combinatorial skills. We find evidence that an experimental-led learning method enhances the introduction of hypothesis testing. That means that students can learn how to evaluate hypotheses within a research-based environment. Class observations show that reliable mathematical tools can back the empirical approach even if the background is not fully clear to students. The approach has a positive impact on students' skills related to their combinatorial and probabilistic thinking. Not only do combinatorial skills and a proper probabilistic understanding help to build appropriate conceptions of the procedure of hypothesis testing, the task of statistically evaluating a

hypothesis supports an understanding of probability. An extension of a combinatorial meaning of probability to a frequentist probability and a kind of abstract risk index seems natural within such an approach. Teachers expressed their dedication to include the methods in their future teaching practice, regardless of whether statistical inference will become part of the curriculum. This opens further questions and possibilities in terms of how the introduction of more complex phenomena, building on already acquired knowledge, may deepen understanding of the "prerequisite" content. The present project demonstrated the feasibility of introducing statistical inference into the Hungarian curricula. The project got the cooperation of the teachers for the reform. The natural links between combinatorics and hypothesis testing will also improve the understanding of probability.

REFERENCES

Borovcnik, M. (2017). *Informal inference—Some thoughts to reconsider* [Paper presentation]. 61st World Statistics Congress of the International Statistical Institute, Marrakech, Morocco.

Bunt, L. N. H. (1967). Probability and statistical inference in the secondary school. *Dialectica*, *21*(1–4), 366–382. https://doi.org/10.1111/j.1746-8361.1967.tb00582.x

Fanshawe, T. (2021). Discovering experimental design: An interactive teaching exercise using Fisher's tea-tasting experiment. *Teaching Statistics, 43*(3), 140–145. https://doi.org/10.1111/test.12287

Fejes-Tóth, P. (2020). Inferential statistics—Research on the introduction of a new topic in the Hungarian high-school curriculum. In G. Ambrus, J. Sjuts, Ö. Vancsó, & É. Vásárhelyi (Eds.), *Komplexer Mathematikunterricht: Die Ideen von Tamás Varga in aktueller Sicht* (pp. 157–180). WTM Verlag für Wissenschafliche Texte und Medien. https://doi.org/10.37626/GA9783959871648.0.10

Fisher, R. A. (1935). *The design of experiments.* Oliver & Boyd.

Garfield, J. B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). Learning to reason about statistical inference. In J. B. Garfield, & D. Ben-Zvi (Eds.), *Developing students' statistical reasoning* (pp. 261–288). Springer. https://doi.org/10.1007/978-1-4020-8383-9_13

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM–Mathematics Education, 44*(7), 883–898. https://doi.org/10.1007/s11858-012-0447-5

Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In R. Cohen Kadosh, & A. Dowker (Eds.), *Oxford handbook of numerical cognition* (pp.1118–1134). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.014

Rossman, A. J., & Chance, B. L. (1999). Teaching the reasoning of statistical inference: A "top ten" list. *The College Mathematics Journal, 30*(4), 297–305. https://doi.org/10.1080/07468342.1999.11974074

Suurtamm, C., Thompson, D. R., Kim, R. Y., Moreno, L. D., Sayac, N., Schukajlow, S., Silver, E., Ufer, S., & Vos, P. (2016). Assessment in mathematics education. In C. Suurtamm, D. R. Thompson, R. Y. Kim, L. D. Moreno, N. Sayac, S. Schukajlow, E. Silver, S. Ufer, & P. Vos (Eds.), *Assessment in mathematics education. Large-scale assessment and classroom assessment* (pp. 1–38). Springer. https://doi.org/10.1007/978-3-319-32394-7_1

Vancsó, Ö, Beregszászi, E., Burian, H., Emese, G., Stettner, E., & Szitányi J. (2018). Complex mathematics education in the 21st century: Improving combinatorial thinking based on Tamás Varga's heritage and recent research results. In E. W. Hart & J. Sandefur (Eds.), *Teaching and learning discrete mathematics worldwide: Curriculum and research* (pp. 111–134). Springer. https://doi.org/10.1007/978-3-319-70308-4_8

Varga, T. (1970). Probability through games. A sample of three games. In International Commission of Mathematical Instruction (Ed.), *New trends in mathematics teaching* (Vol. 2, pp. 424–440). United Nations Educational, Scientific, and Cultural Organizations.

Varga, T. (1972). Logic and probability in the lower grades. *Educational Studies in Mathematics, 4*(3), 346–357. https://doi.org/10.1007/BF00302583

Varga, T. (1983). Statistics in the curriculum for everybody—How young children and how their teachers react. In D. R. Grey, P. Holmes, V. Barnett, & G. M Constable (Eds.), *Proceedings of the First International Conference on Teaching Statistics* (Vol. 1, pp. 71–80). Teaching Statistics Trust. https://iase-web.org/Conference_Proceedings.php?p=ICOTS_1_1982