# A complex concept about statistical inference and a planned school experiment based on it

Ödön Vancsó[1]; Manfred Borovcnik[2]; Peter Fejes-Tóth[3]

[1] Eötvös-Lóránd University, Centre for Didactics in Mathematics, Budapest, vancso.odon@ttk.elte.hu
[2] University of Klagenfurt, Institute of Statistics, Klagenfurt
[3] Szent-István University, Institute for Biometry and Agrarian Informatics, Budapest

**Abstract:**
Carranza and Kuzniak (2008) analysed the negative effects of reducing probability to a purely frequentist notion on the students' perception of the methods to learn. Probability is incomplete and cannot be fully understood unless it is linked to statistical inference. Statistical inference usually is reduced to very simple methods and, more recently, attempts have increased to follow the ideas of informal inference, which reduces it even to a mere significance test based on simulation and a narrow interpretation of probability as relative frequencies. This informal approach is favoured by its simplicity. Nevertheless, there still is a need for innovative studies on teaching and learning probability and statistical inference from a broader perspective, which would also make it possible to integrate Bayesian methods into inference, partly based on the Bayesian formula. We develop the different views (Bayesian and classical inference) in parallel – so that students can understand either of them better. To support this challenging task, we integrate philosophical considerations, use paradoxes, and make extensive use of computers. We build on encouraging feedback from our courses to date. In Hungary, we are currently in the process of reforming the curriculum, in which we will try to find ways to implement our ideas sustainably, especially among in-service teachers. Of course, our reflections will also be revised with regard to the feedback of the experimental teachers involved.

**Key words:**
Inferential statistics; Bayes statistics; Curriculum; Final exam; Complex mathematics education; Hypothesis test

**Introduction:**

In Hungary, inference statistics is not taught in secondary schools. This paper describes the didactic concept of an experiment with elaborated pilot teaching material in some schools and the results of this experiment with regard to the pandemic circumstances. The concept is based on a parallel introduction of classical and Bayesian inference and therefore uses a rich concept of probability and can be characterised by an informal approach, see Borovcnik (2020). Our aim is to formulate a local part of the higher-mathematics curriculum in Hungary, including the requirements of the final examination, and to develop useful teaching material on the subject. We will report about the pilot experiment.

The history of teaching statistical inference in Hungarian schools began decades earlier. T. Varga was the first to introduce probability, statistics, and combinatorics into the curriculum of his Complex Mathematics Education Experiment in 1978, which regulated primary education from Grade 1 to 8 (6-14 years) in Hungary. One of his main ideas is to introduce modern mathematical terms and topics into school mathematics as early as possible. *Let's play mathematics* (Varga 1972a) is about playing games, in which "chance" has an important component for children aged 9 and over. Further examples have been collected in Varga

(1970), and a scientific analysis of teaching the subject in Hungary can be found in Varga (1972b). As a result of his efforts, a new secondary-school curriculum was introduced in 1979, which included statistics *only in higher-level* mathematics (combinatorics and probability played a key role, but only combinatorics tasks were included in the final examination, none of the tasks related to probability). As a result of Varga's conceptual failures, statistics was not established as an integral part of school mathematics until 2005, although some elements of statistics had to be prescribed in the curriculum, this topic occurred very rarely in school practice and in textbooks. The relatively low impact of the new curriculum is due to Varga's early death in 1987 and to the lack of sufficiently staffed teacher training courses to prepare the primary school teachers for the new ideas of complex mathematics teaching; the original curriculum was revised in 1986. Nevertheless, T. Nemetz continued his experiments on Varga's concept for the secondary school in the 1980s and wrote textbook chapters on probability and statistics combining the two fields much better than in many other sources (see Nemetz & Kusolitsch 1999). Nemetz supervised Vancsó's dissertation entitled "*Classical and Bayesian Statistics from Didactic Perspective*", but this was already done 15 years later (Vancsó 2006). In 2005, a new final examination was introduced in Hungary, which represented a major step forward in the teaching of probability and statistics; since education in Hungary is highly results-oriented, the new final examination has had a significant impact on daily school practice. However, the examination only covered tasks from descriptive statistics and probability with tasks related to inferential statistics still missing.

To prepare for this new examination, a group was formed, in which Vancsó represented the universities that were aware of the importance of his own research field of probability and statistics in mathematics education. Finally, the demands on statistics were limited to descriptive statistics, because the original idea with inference statistics ultimately fell out of the considerations, mainly due to the decision that mathematics as school subject in Grades 11 and 12 lost one lesson per week. Despite all this, a new course on Bayes statistics was started in 2006 to 2008 in teacher training based on earlier theoretical research (Vancsó 2006) treating Bayes methods in parallel with the classical point of view. The outcome of these courses was further analysed to revise the material and the examination papers and published in Vancsó (2009), among others. Our research experiment aims to change the current situation and to integrate inferential statistics into the curriculum and the final examinations, at least on higher levels. In order to implement this plan, we have set up a working group within the research group "Complex Mathematics Education" under the auspices of the Hungarian Academy of Sciences (HAS) and Eötvös-Loránd University (ELTE). The HAS-supported project started in 2016 at and will be completed this year; the present authors are involved. The aim was to explore the legacy of Tamás Varga, who developed a new concept for school-mathematics (Complex Mathematics Education) in the 1960s and 70s.

A summary of the project's result was reported at the Varga 100 International Conference 2019 in Budapest. The lectures were published in two special issues of *Teaching Mathematics and Computer Science* (TMCS 2020, 18(3), (4)) and in a book (in German); see Ambrus et al (2020). Several speakers addressed topics in stochastics, which was one of Varga's favourite topics (see, e.g., Borovcnik 2020). We intend to continue the HAS-ELTE project, in which we planned a larger experiment in schools with students of higher-level mathematics. The idea is to develop new learning paths in hypothesis testing in different ways. We prepared materials for teachers that they could use in class for about 6-8 lessons. The first ideas and experiments were formulated and carried out at four schools in 2019-20. However, the pandemics prevented a prolongation of the experiment. At this year's ICME-14 in Shanghai, P. Fejes-Tóth will present the concept of the experiment and a report on the first results of testing the material in class. The background ideas can be taken up in Borovcnik et al (2020) in German and in Fejes-Tóth (2020) in English.

This summer semester (2021), Vancsó gives a seminar for doctoral students on Bayes statistics. One student understood why she had not yet seen the correct concept of confidence interval and its correct interpretation. We have briefly summarised the course concept, which also incorporates ideas from Gigerenzer (1993), who also described a hybrid form of classical and Bayesian statistics and their interpretation in connection with psychological processes in our minds. The philosophical background to Bayesian statistics and the interpretation of classical methods can be found in Wickmann (2001) and further papers published in Borovcnik, Engel, & Wickmann (2001), a book entitled *Suggestions for Teaching Stochastics* (Anregungen zum Stochastikunterricht). Using Vancsó's formulation from his course, "we are interested to find such an interval that covers with high probability (e.g., 0.95 or 0.99) the unknown parameter that we have to estimate". The fact is that the classical confidence interval has no such a property. All that means is this: In the long run (repeat to obtain samples and calculate the confidence interval), we get an experimental relative frequency of 0.95 (0.99) that the "true" value of the unknown parameter belongs to the confidence interval. However, there is a slight difference when you say, "On average, we get an interval that contains the unknown parameter with probability 0.95", or when you state, "The true parameter is in this concrete interval that we just got".

Only Bayesian thinking can answer our original question, which interests us, but we have to pay for it. Indeed, we must acknowledge or accept subjective probability and abandon the idea of "objectivity". Gigerenzer's paper described the differences between the desired interpretation and the correct one. Our experience in teacher training supports this view, see Vancsó (2009). Borovcnik (2020) summarised the main aspects of the subject in his plenary lecture at Varga 100 in Budapest. He also mentioned our initial experiment and highlighted the new idea: "Informal pathways can create appropriate learning paths to show the nature of statistical inference. One may also learn from the paradigmatic examples of Batanero and Borovcnik (2016) and Borovcnik et al (2020). These examples are about playful activities that are in line with Varga's intention to provide challenging situations to learners and let them explore the situations."

After the introduction, we begin to write about the experiment, which is the main objective of the present paper. In the preliminary process of introducing hypothesis testing in the Hungarian secondary school-mathematics curricula, we test the effectiveness and feasibility of the designed teaching material. A pilot teaching of this larger-scale experiment was carried out in four classes in three secondary schools, with a total of four teachers. The final number of students participating in the pilot scheme was 65. The curriculum consists of six lessons of 45 minutes. We provided the participating teachers with teaching materials, lesson plans, presentations for each lesson, and simulations in Excel for throwing dice and coins.

**Content of the curriculum:**

In the six teaching units, students are guided through the entire process of hypothesis testing, from simple experiments such as throwing the dice to understanding complex mathematical phenomena. In three experiments, we introduced basic concepts of hypothesis testing. The first is the "Lady tasting tea" experiment with water, in which it is necessary to decide whether the water is mineral or plain tap water. The questions asked here lead towards hypothesis testing, but the calculations that students have to perform for decision-making are based on simple combinatorial knowledge that they have already acquired. The second experiment is about tossing a coin, in which the students have to decide whether the coin is loaded or not. In this experiment, we presented two different methods of decision-making. On the one hand, the decision can be made by a purely combinatorial method based on probabilities. On the other hand, however, the method of using the chi-squared test can be introduced. In the third experiment, the students roll the dice and have to decide whether the die is loaded or not.

For this purpose, fair and loaded dice were produced with a 3D printer, which are similar in appearance. Students first examined the dice physically and tried to make a first assumption. Then, based on a frequency table of 300 throws, they make a subjective decision, then, in order to arrive at a mathematically substantiated decision, they perform a chi-squared test. In this step, Bayesian ideas could be introduced, in which we use our prior knowledge of the situation. In our case, we make our own hypothesis about the regularity of the dice after examining it. With the Bayes theorem, it is possible to talk about prior and posterior distributions. After discussions with teachers involved, we decided that they should offer the students some Bayesian material and not to go deeper into this topic within this experiment.

The aim of the curriculum is multifaceted, ranging from deepening the understanding of combinatorial methods to introducing the logic, techniques and crucial concepts of hypothesis testing (like Type-I and Type-II errors). The general goal is helping students understand how probability interferes in real life work, how we can make decisions under uncertain circumstances, and how is it possible that even a good and mathematically sound decision can turn out to be bad without anyone making a mistake. Closely related to this, the proper way of interpretation of the results of hypothesis testing was a key point in the curriculum. The understanding that in most cases no definitive answer can be given and that we can only speak of more or less probable states of reality is at the core of this method, but often proves difficult for students to understand.

**Methodology:**

With the pilot test, we wanted to show whether the designed curriculum seems appropriate and whether the measurement methods seem to provide reliable information about this adequacy. We wanted to test the feasibility, acceptance and effectiveness of the course design among students and teachers.

In order to achieve this objective, we had to assess the attitudes and perceptions of students and teachers regarding the design of the material and the lessons, as well as students' actual understanding, i.e., the learning outcomes. To this end, we have designed a study with mixed methods and with repeated data collection. We use different methods of formative assessment (Suurtamm et al. 2016) as well as two formal post-tests after the end of the programme with different time delays. Most assessment techniques were quantitative, with the exception of a semi-structured interview with the teachers on their experiences, perception and evaluation of the curriculum. The quantitative tools used were a formative assessment in a pre-post-test design, and an attitude-assessment questionnaire distributed to the students. With regard to the pre-post-test in the different knowledge areas, different conditions of a successful transfer could have been applied.

In case of knowledge in combinatorics and probability, where students had a measurable knowledge prior the program, we can consider the knowledge transfer successful if there was a significant increase in correct answers. In case of the new material, which was hypothesis testing itself, we measured the success of the knowledge transfer in terms of correct solutions in the post test. In the attitude-assessment questionnaire, the questions split into two categories: (1) feasibility, (2) acceptability of the teaching. Dimension (2) breaks into the sub dimensions of (a) the general attractiveness ("how much did you like it") and (b) the perceived usefulness of the course and the material. The students had to rate 20 statements regarding the features of the course or the material in general; we used a Likert scale ranging from 1 to 6 to express how much they agree or disagree with the statements.

**Results:**

Regarding the results of the pilot experiment, the curriculum proved to be suitable in all three main performance dimensions. Based on the comparison of pre- and post-test, students' achievement was significantly better in the post-test regarding combinatorial questions on sampling both with and without replacement (see Table 1).

Table 1. Answers to the combinatorial problem in sampling with replacement in pre- and post-test

| *Classification of answer* | *Pre-test* | *Post-test* | |
|---|---|---|---|
| Completely wrong answer | 32 | 2 | Chi-squared = 28.9, df = 2, p value < $10^{-6}$ |
| Partly correct answer | 13 | 12 | |
| Correct answer | 6 | 30 | |

In the pre-test, we could not yet include tasks on hypothesis testing. That is why we invented the "umbrella" task (where students had to decide to bring or not to bring an umbrella to a hiking tour). There were several questions here: Which decision do students make and which aspects do they consider in a situation with uncertain outcome? Which kind of "expected value" do they base their optimisation on? Do they take into account the "cost" of unnecessarily carrying the umbrella, or soaking if leaving the umbrella at home? What do they think about the quality of the weather forecast?

In the evaluation of the umbrella task, we gave a score of 0 or 1, depending on whether their explanation was fitting. An interesting recurring mistake was mentioning a probability greater than 50% in the argument, e.g., "It was a very good decision because it will be more than 50% chance that it would rain". This kind of reasoning does not reflect to the "cost" of possible bad decisions, and the weather forecast is also misinterpreted, as if a more than 50% chance of raining would mean raining for sure, while a less than 50% chance would mean not raining for sure. Altogether, 76% of students provided an acceptable reasoning.

In the post test, a specific kind of hypothesis test (chi-squared goodness-of-fit test) had to be performed. During the evaluation of the task 5 points where examined:

1. Determination of the calculated value (also called test statistics, chi-squared value)
2. Decision-making with 0.05 Type-I error level
3. Determination of possible Type-I error
4. Decision-making with 0.01 Type-I error level
5. Determination of the possible Type-II error

91% of the students calculated the chi-squared value correctly. No calculation error was made: the remaining 9% was not correct because students did not even start the calculations. Regarding points 2 and 4: those who came to the right decision at 0.05 level, succeeded on 0.01 level, too. 78.2% of students came to a right decision on both significance levels. around 90% of those who calculated the chi-squared value correctly, also made the decision correctly.

In textual response interpreting the results, however, the students were succinct and often used incorrect wording, making definitive statements instead of referring to probability. They wrote claims like "*data is out of date*" or "*data is not out of date*". One of the most important lesson related to hypothesis testing is that one cannot state anything with certainty. The correct phrasing would be "*based on the values obtained, we have good reason to doubt the accuracy of the data at a significance level of 0.05 percent*" or "*the values obtained are not convincing enough to reject our initial assumption at a significance level of 0.01 percent*".

Regarding students' and teachers' attitudes towards the curriculum, the answers verified both feasibility and acceptability. The majority of students stated that they found the classes understandable, they think that the topic is of importance, and they overall liked the program. From teachers' point of view, the curriculum proved to be also feasible, however some of them noted that they felt somewhat insecure needing to teach something without the stable knowledge regarding the underlying mathematical theory.

**Discussion and future plans:**

The results of pilot project are encouraging (underpinned by the positive change in combinatorial problem solving after the experiment) and provide us a basis for argumentation to introduce inferential statistics into the curriculum and the final examination. Before this process begins, we still need to analyse the interviews with participants of the pilot project and prepare an experiment with more schools (including more teachers and students) involved. We are trying to revise and expand the Bayesian part of the pilot material even more precisely, because this topic was partly neglected by the teachers due to their unfamiliarity with the ideas and to computational problems. It was too far from their mathematical content knowledge. We are working together with A. Krajcsi, a psychologist who developed a software for teaching statistics to psychology students at Eötvös-Loránd University, named by CogStat, (www.cogstat.org/download.html). His experiences with this software with university students are encouraging especially as psychology students are on average less inclined to mathematics, so we are convinced that this tool can also be used by secondary-school students. It seems possible to set up an expanded and revised experiment in the school year 2022/23, in which CogStat will be adapted and used for secondary-school students. Following the experiments, a detailed proposal for the curriculum for the decision makers is planned for autumn 2023. The new topic could be introduced in the school year 2024/25. Our plan provides ample time for discussion and preparation of teachers (in-service teacher-training courses) on both types of inferential statistics. An important lesson from the introduction of Varga's new curriculum in the 1980s is that teachers need to be better prepared. We are trying not to repeat this failure of the past.

**Acknowledgement:**

The authors are grateful that the *Complex Mathematics Education Research Project* hosted by MTA-ELTE (HAS-ELTE) has financially supported the pilot experiment.

**References:**

Ambrus, G., Sjuts, J, Vancsó, Ö., & Vásárhelyi, É. (Eds.) (2020). *Komplexer Mathematik-unterricht: Die Ideen von Tamás Varga in aktueller Sicht*. Münster: WTM Verlag.

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

Borovcnik, M. (2020). Many paths lead to statistical inference – Should teaching it focus on elementary approaches or reflect this multiplicity? *Teaching Mathematics and Computer Science, 18*(4), 259-294.

Borovcnik, M., Fejes-Tóth, P., Jánvári, Z., & Vancsó, Ö. (2020). Experimente zur Einführung von Ideen und Denkweisen statistischer Inferenz im Gymnasium. *Stochastik in der Schule, 40*(1), 18-27.

Carranza, P., & Kuzniak, A. (2008). Duality of probability and statistics teaching in French education. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching statistics in school mathematics. Challenges for teaching and teacher education*.

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). Münster: WTM Verlag.

Gigerenzer, G. (1993). The superego, the ego and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological issues* (pp. 311-339). Hillsdale, NY: Lawrence Erlbaum.

Halmos, M. & Varga, T. (1978). Change in mathematics education since the late 1950's ideas and realisation in Hungary. *Educational Studies in Mathematics, 9*, 225-244.

Nemetz, T., & Kusolitsch, N. (1999). *Guide to the empire of random*. Budapest: TypoTex.

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*. Cham: Springer.

Vancsó, Ö. (2006). *Classical and Bayesian statistics from didactic perspective*. Dissertation: University of Debrecen.

Vancsó, Ö. (2009). Parallel discussion of classical and Bayesian ways as an introduction to statistical inference. *Intern. Electronic Journal of Mathematics Education, 4*(3), 291-322.

Varga, T. (1970). Probability through games. A sample of three games. In ICMI (Ed.), *New trends in mathematics teaching II* (pp. 424-440). Paris: UNESCO.

Varga, T. (1972a*). Let's play maths! I-II. Book for children age 9 and up*. Budapest: Móra Kiadó

Varga, T. (1972b). Logic and probability in the lower grades. *Educational Studies in Mathematics 4*(3), 346-357.

Varga T. (1978). *Curriculum of Complex Mathematics Education*. Budapest: Ministry of Education.

Varga, T. (1983). Statistics in the curriculum for everybody – How young children and how their teachers react. In D.R. Grey et al. (Eds.), *Proceedings of the First International Conference on Teaching Statistics, Vol. 1* (pp. 71-80). Sheffield: Teaching Statistics Trust.

Wickmann, D. (2001). Inferenzstatistik ohne Signifikanztest In: M. Borovcnik, J.; Engel, & D.; Wickmann (Eds.), *Anregungen zum Stochastikunterricht* (pp. 133-138). Hildesheim, Berlin: Franzbecker