INTRODUCTION OF HYPOTHESIS TESTING THROUGH AN INFORMAL INFERENCE APPROACH IN A MEXICAN HIGH SCHOOL

Eleazar Silvestre Castro, Victor N. Garcia & Ernesto Sanchez
CINVESTAV-IPN, Mexico
eleazar.silvestre@gmail.com

We report the main results of a series of design experiments in which two full high school groups (16-18 y.o.) faced hypothesis testing using random simulations via Fathom; one group (N=36) had no previous contact with statistical inference, while the other (N=42) was already engaged in the study of sampling distributions through an informal approach. Despite their background differences, we observed some common and similar patterns in students' reasoning during the experiments in relation to both achievements and difficulties in their performance: conflicts in the identification of the rejection zone and computation of the p-value; a tendency to assume the test as a proof of truth; and a gradual incremental and explicit usage of key concepts.

INTRODUCTION

Hypothesis testing (*HT*) is a fundamental concept in introductory inferential statistics, and at the same time, there's a considerable amount of evidence that indicates a diverse variety of errors and misconceptions that a wide spectrum of individuals tends to present in relation to its uses and understanding (e.g., Castro-Sotos, Vanhoof, Van den Noorgate, & Onghena, 2007). These difficulties have been linked to (a mix of) a number of reasons, such as the concept's great epistemological complexity (Batanero, 2000), and the prevalence of traditional teaching practices that prone to reproduce the same mistakes that teachers' commit (Liu & Thompson, 2009; Harradine, Batanero & Rossman, 2011). In recent decades, efforts made by the movement called *informal inferential reasoning* (e.g., Zieffler, Garfield, delMas & Reading, 2008) derived in some didactic proposals that heavily rely in the use of *simulations* to construct empirical *sampling distributions* (*SD*'s) to deal with some basic and most troublesome elements of the inference curricula; in the case of HT, these resources have been used to promote a more *frequentist/stochastic approach* to introduce some of the main ideas at the concepts' core (Rossman, 2008; Batanero & Diaz, 2015).

As well as other developing countries, since Mexico's curricula has suffered recent modifications that could translate into a bigger incorporation and a greater emphasis in stochastic content, coupled with the use of (educational) statistical software at a younger age (Cuevas, 2012) – an already elaborated scenario that pinpoints some related and complex phenomena on itself; for example, the understanding of students' learning process about sampling in a technology rich environment (Lipson, 2002) –, a more urgent need for evidence that show how students' reasoning behaves in the context of these tools and content arises, especially if such a many-sided concept like HT has been already and increasingly introduced in secondary education over the recent years; such information becomes an even more a key variable to form solid bases for successful planning and execution of statistical learning processes. Thus, in this report we focus on the main research questions: a) *what type of reasoning, in terms of strategies and arguments, high school students present when they first face HT in an informal approach?* b) *What difficulties in students' learning process around HT emerge in an informal approach?* We aim to contribute to these interrogations by reviewing two high school Mexican's experiences with HT using *Fathom* in two different settings; one participant group of students had no previous contact with random simulations nor statistical inference at the moment of facing HT, while the other one previously attended sampling and empirical sampling distributions tasks that also included the *notion of the most likely/typical values*; the purpose of this review is to identify patterns in students' reasoning during each experiment in relation to both achievements and difficulties in their performance. It is important to note that the results of each study cannot be compared due to some methodological differences.

METHOD

Background and main features of the experiments. Our results stem from the retrospective analysis of two parallel but closely related studies executed during 2015 and 2016, S_A and S_B (Silvestre &

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

Sanchez, 2016; Sanchez, García-Ríos & Mercado, 2017), that share the general objective of gaining a better understanding of how to introduce and develop reasoning with HT in the Mexican high school; both studies share key methodological aspects such as following the *design experiments scheme* (Cobb, Confrey, diSessa, Lehrer & Schauble; 2003), the common vision of an *informal inferential reasoning* based on Zieffler et. al's conception (2008) complemented with an *inferentialist approach* (Bakker & Derry, 2008), and the use of Fathom coupled with the *learning trajectories* mechanism to plan and structure the activities. The main differences between these studies derive from the trajectories' content and participants' background status at the moment of dealing with HT: participants in S_A faced problems related only to HT having no previous contact with statistical inference but only mild interactions with descriptive statistics during some preceding mandatory courses; instead, participants of S_B initially engaged in the study of SD's using binomial models and random simulations to explore base-urn model situations before facing HT (at the end of their trajectory). Among some of the important notions attended in S_B's trajectory, such as the law of large numbers or the effects of sample size in the SD, we also included the notion of the most/least typical/likely values in two related formats: those statistics that present the largest/smallest frequencies, or those whose frequencies englobe the most/least of all samples at a pre-fixed level (such as 90% or 95% of all values around SD's center; Figure 1, below):

Figure 1. Fathom simulations of the SD used in S_A tasks (above); an empirical SD with the "most likely/typical values" range at 95% used in S_B tasks (below)

Therefore, since S_B participants had an "extra advantage" of three weeks class sessions that directly addressed the aforementioned notions of sampling and estimation via the SD, we envisioned that these students might be more capable to deal with' and grip more elements embedded in the HT's structure; so, problems used in S_A were designed to follow Fisher's conception based on the assumption for it to be a more natural-logic reasoned pathway for students

to grasp (García & Sanchez, 2014; Batanero & Diaz, 2015), and the one included in S_B follows Neyman & Pearson's, assuming that this version requires a more expanded and robust statistical machinery (the managing of two hypotheses, concepts of significance level, acceptance/rejection zones and p-values, errors I and II, etc.). Despite this methodological difference in each study we are interested in the students' performance within the informal approach that S_A and S_B proposed.

Participants, data and analysis tools. Two full high school groups participated in the studies; S_A consisted of a *regular* group of 36 students enrolled in their 2nd year of high school (*16-17 y.o.*), and S_B consisted of another regular group of 42 students enrolled in their 3rd year (*17-18 y.o.*). Collected data is mainly composed by students' responses to the activities, and complemented with field notes, memos and non-structured short interviews; we used *Grounded Theory* basic coding procedures (Birks & Mills, 2011), and *SOLO Taxonomy* (Biggs & Collis; 1982, 1991) to measure and estimate students' reasoning development during the trajectory, in relation to their ability to identify and properly use key aspects in each situation of the trajectory (e.g. estimating and using a p-value out of an empirical SD; accepting/rejecting the appropriate hypothesis).

Instruments and execution. Activities of both trajectories part from non-mathematical contexts. Since problems in S_A align more with Fisher's conception, problems require students to focus on the evaluation of sample data's strength, in order to *accept/reject a (null) hypothesis* (example A); since problems used in S_B align more with Neyman & Pearson's conception, the emphasis here is to *choose between one of two hypotheses* (example B):

- Example A / Activity 1: "Coca-Cola's advertising campaign *assures* that most of the population (more than 50%) prefers its product over Pepsi's. To corroborate this, an experiment was conducted where 60 persons randomly chosen tried both beverages in a blind test; 35 of the participants chose the Coke option. *Is the hypothesis 'more than 50% prefers Coca-Cola over Pepsi' correct*?" – [H_0: $P \leq .5$ is accepted at $\alpha = .05$]

- Example B / Activity 6: "A company possesses four industrial-level machineries dedicated to manufacture laptop motherboards. Inevitably, the machines start to produce defective cards after some period of time; the Quality and Optimization Department declared that *deficient cards are allowed in no more than 10% of the overall production*, otherwise the corresponding machine must be ceased and sent to an *inspection* for *possible repair*. After a certain day of regular production, a random sample of 120 cards was taken of each machine, obtaining the results of M_A: 42 defective cards, M_B: 21, M_C: 27 and M_D: 15; *which machine (s) do you consider should be retired for inspection*?" – [H_0: $P \leq .1$ is accepted only for M_D at $\alpha = .025$]

Regarding execution processes, students were randomly arranged in pairs in each class session. After posing the problem, a small dialogue was triggered to assure participants grasped some key aspects of' (e.g., where one might find random processes during the experiments) and what the situation solicited. Then students engaged in a free and collaborative team work using and manipulating *Fathom* to simulate empirical distributions (accumulating from 300 to 500 samples) of the (null) hypothesis, and generated final reports where they explained both their arguments and reasoning for their solutions to the tasks; the teacher/researcher mostly helped participants to induce central questions for inquiry and overcome technical difficulties. A more detailed version of the activities can be consulted in Silvestre & Sanchez (2016), García & Sanchez (2017), and Sanchez et. al. (2017).

RESULTS

The next results (Table 1) come from students' first experiences with HT, where they faced the aforementioned problems; codes appear in descending order according to their frequencies, and $CX_{A,B}$ is used to denote one pair's response. The vast majority of participants in both groups initially *confirmed* both hypotheses (78% in S_A and 66% in S_B), arguing that since the given statistic exceeded *P* the hypothesis was correct; for example, $C3_A$ mentioned "we can call the majority from 51% on, and the experiment's result showed that 35 out 60 people preferred Coca-Cola which equals 59% of the total [(59% x 60) = 35.4]. We can conclude that they are not wrong with their guess"; $C8_B$ responded "we would send all machines for *repair*, since the limit of defective cards is 10% and they all exceeded that percentage, the maximum should be 12 in a

sample of 120". Students seemed unable to identify the presence of random process and sample variability within the context of each situation; while no pair in S_B proposed to simulate any sample or distribution, only one in S_A proposed to simulate a binomial distribution with $n = 60$ and $P = p_0$ (taking the statistic as the parameter of the simulated distribution). After the teacher/researcher re-introduced the ideas of sample variability and sampling distribution by posing questions such as "*If we had a population with P = .5, how weird or typical might these results be?* / *If we already know that Machine Z (M_z) has a 10% production of defective motherboards, how would a distribution of samples of n=120 each taken from there look like*?", students were guided to simulate the corresponding SD and asked to reformulate their previous decision having the simulated data at their disposition.

Table 1: Codes for students' responses in each problem

HT Components	*Sub-categories*	
	Example A	*Example B*
Hypotheses and simulation	** Lack of sampling variability -- Statistic as parameter	
P-value and zones	-- Located in the "majority zone" -- Mode of the SD -- Pseudo p-value	-- Most/least likely values -- Pseudo p-value -- Heuristic complement
Conclusion (decision making)	** Mostly correct/appropriate ** Inversed or doubtful	
Acknowledgment of uncertainty	-- Pop. Distribution instead of SD -- (Technical) certainty	-- "Repair" instead of "inspect" -- Probability in language

In light of the SD, students' arguments and reasoning were substantially different in each case; the majority of participants in S_A (75%) made their decision by dividing the SD in two equal parts (samples with less/more than 50%) and locating where the statistic fell; others compared the mode of the SD to its center (15%): *C13$_A$* justified "our largest value in the simulation was 29 out of 60 that liked Coca-Cola the most, then we can see that less than 50% prefer Coca-Cola". Since students in S_B had already been working with the notion of most likely/typical values in a general sense (without discriminating low/high typical sample values according to regions), they roughly reasoned in a similar way to a two-tail test; most of them (80%) formed and used the regions around the SD's center that usually included 95% of all samples and identify where the statistic fell: "We would send machines A, B and C for *repair* because they are outside of the most likely values range (they represent 2.5% of all samples). Machine D would be the only one not to be sent, because its % lies within the other 95% of all samples" (*C10$_B$*). Only a few of these students (19 %) focused or made emphasis in the upper region of the SD: *C3$_B$* responded "We are *certain* that we would send machines A, B and C for repair because they present the most defective cards and exceed the limit of the most likely/typical value (the range of 10-15%), so machine D *has less probability* in producing defective cards". Despite using these types of arguments, participants of S_B such as the former often recurred to some heuristic-like reasoning by adding "…and because machine A/B/C presents the most defective cards". A common strategy in both groups was the usage of the samples' frequencies to form a *pseudo p-value*: they counted the number of samples of the given statistic in the simulated SD, without considering more extreme values, to evaluate how likely or plausible the result might be; this strategy was utilized by a larger set of S_A students' than S_B's probably because of the more "inexperienced" participants in S_A.

Although most students appropriately rejected the corresponding hypothesis in each group, a small fraction inverted the final decision, or abstained from doing so by demanding more tests (samples) or to increment sample size. Hence, when interrogating students about their confidence in their final decision, participants in S_A exhibited that they believed the simulated SD directly represented the population's distribution, in later moments, they also added that "*following the procedure correctly*" made them feel more secure about their conclusions. In a subtler manner, participants in S_B quickly changed their discourse by substituting "*inspect*" for "*repair*"; they later mentioned they believed the test implied proving whether the machine had a malfunction or not. Only a very small sub-group of students attempted to moderate this confidence by integrating probability statements about the machines' production (such as *C3$_B$*'s response).

As the experimentation with the learning trajectories in both groups continued, a notorious change in students' discourse became noticeable, they began to explicitly recall and integrate more statistical objects and concepts in their responses and justifications. Thus, we considered appropriate to estimate students' reasoning development; the next graphs (Figure 2) represent this cognitive progress in terms of SOLO levels on each study:

Figure 2. SOLO outcomes in each trajectory

Even though each reasoning level is built independently and with different components, the distribution of outcomes in both studies is roughly similar, the pattern in students' responses suggests an improvement in their learning as the trajectory unfolded in both experiments; however, two great differences can be noticed: (i) the relational level was more achievable than the extended-abstract for participants of S_B in their first experience with the HT (at the end of their learning trajectory), and (ii) even with just a few more experiences than S_B participants, two thirds of S_A's students were able to reach the relational or abstract level at the end of their trajectory. In the first case, we suspect that student's recent practices of sampling mostly rid them from the main technical-procedural aspect of the test process by quickly focusing and using resources already grounded in the SD. In the former case, although much of students' reasoning refined in terms of a considerable extent of the HT's structure [(1) associating the HT procedure with a particular SD, (2) the technical ability to simulate the corresponding SD, (3) the computation of the p-value and acceptance/rejection zones, and (4) the appropriate final decision making], we observed that two of the most difficult aspects for students to overcome was to correctly identify and justify the distribution of the hypotheses at the beginning of the task, as well as the recurrence to elaborate final reports that didn't reflect any degree of uncertainty and often mixed with the idea of relating the achievement of "confidence" to procedural execution.

CONCLUSIONS AND DISCUSSION

In relation to question a) we observed that students' reasoning presented great differences mainly linked to the understanding and uses of the SD when attempting to elaborate a more rational procedure for the HT's final decision, like using more limited strategies to determine an acceptance/rejection zone in the case of the more "inexperienced" participants. On the other hand, both groups presented close similarities at both start and at the end of the activity, where patterns as the neglection of sampling variability's presence and the tendency to assume the test as proof of truth emerged. In relation to question b) although we observed that most of our students' reasoning development became more robust and better oriented by incorporating and relating more elements as the trajectory unfolded (Rossman, 2008; Batanero & Diaz, 2015) –maybe as a product of a didactic treatment that privileged a more holistic and inferential use of the statistical objects involved (Bakker & Derry, 2008)– the recurrent presence of the aforementioned difficulties suggests a greater difficulty that lies more directly into the managing of these particular aspects. On such assumption, these patterns should be considered for instruction purposes since overcoming them is required for a more appropriate and deeper understanding of the HT; after all, the proper managing and understanding of the p-value as a conditional probability and the types of errors I and II that are intrinsic to the test are indispensable for the concept's basic scheme.

REFERENCES

Bakker, A. & Derry J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning*. *13*(1-2), 5-26.

Batanero, C. (2000). Controversies around significance tests. *Mathematical Thinking and Learning*, *2*(1-2), 75-98.

Batanero C. & Díaz C. (2015). Aproximación informal al contraste de hipótesis. In J. M. Contreras, C. Batanero, J. D. Godino, G.R. Cañadas, P. Arteaga, E. Molina, M.M. Gea y M.M. López (Eds.), *Didáctica de la Estadística, Probabilidad y Combinatoria, 2* (pp. 207-214). Granada, 2015.

Biggs, J. B., & Collis, K. (1982). *Evaluating the Quality of Learning: the SOLO taxonomy*. New York, Academic Press.

Biggs, J., & Collis, K. (1991). Multimodal learning and the quality of intelligent behavior. In H. Rowe (Ed.), *Intelligence, Reconceptualization and Measurement* (pp. 57-76). New Jersey: Laurence Erlbaum Association.

Birks, M. & Mills, J. (2011). *Grounded theory: A practical guide*. California: Sage.

Castro-Sotos, A. E., Vanhoof, S., Van den 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.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R. & Schauble, L. (2003). The role of design in educational research. *Educational Researcher*, *32*(1), 9-13.

Cuevas, J., H., (2012). Panorama actual de los estándares educativos en estocástica. *Revista digital Matemática, Educación e Internet*, *12*(2). México.

García, V. N. & Sánchez, E. (2014). Razonamiento inferencial informal: el caso de la prueba de significación con estudiantes de bachillerato. In M. T. González, M. Codes, D. Arnau & T. Ortega (Eds.). *Investigación en Educación Matemática XVIII*, (pp. 345-357). Salamanca: SEIEM

García V. N. & Sanchez, E. (2017). Exploring high school students beginning reasoning about significance tests with technology. In Galindo, E., & Newton, J., (Eds.). *Proceedings of the 39th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, (pp. 1032-1039). Indianapolis, IN: Hoosier Association of Mathematics Teacher Educators.

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* (pp. 235- 246). A Joint ICMI/IASE Study.

Liu, Y. & Thompson, P. W. (2009). Mathematics teachers' understandings of proto-hypothesis testing. *Pedagogies*, *4*(2), 126-138.

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics*, Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal*, *7*(2), 5-19.

Sánchez, E., García-Ríos, V.N. y Mercado, M. (2017). Desarrollo del razonamiento sobre pruebas de significación de estudiantes de bachillerato en un ambiente tecnológico. In J.M. Muñoz-Escolano, A. Arnal-Bailera, P. Beltrán-Pellicer, M.L. Callejo y J. Carrillo (Eds.). *Investigación en Educación Matemática XXI*, (pp. 447-456). Zaragoza: SEIEM.

Silvestre, E. & Sanchez, E. (2016). Patrones en el desarrollo del razonamiento inferencial informal: introducción a las pruebas de significancia en el bachillerato. In J. A. Macías, A. Jiménez, J. L. González, M. T. Sánchez, P. Hernández, C. Fernández, F. J. Ruiz, T. Fernández y A. Berciano (Eds.), *Investigación en Educación Matemática XX* (pp. 509-518). Malaga: SEIEM.

Zieffler, A., Garfield, J., delMas, R. & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistical Education Research Journal*, *7*(2), 40– 58.