EXPLORING HIGH SCHOOL STUDENTS BEGINNING REASONING ABOUT SIGNIFICANCE TESTS WITH TECHNOLOGY

Víctor N. García
CINVESTAV-MEXICO
nozairg@hotmail.com

Ernesto Sánchez
CINVESTAV-MEXICO
esanchez0155@gmail.com

In the present study we analyze how students reason about or make inferences given a particular hypothesis testing problem (without having studied formal methods of statistical inference) when using Fathom. They use Fathom to create an empirical sampling distribution through computer simulation. It is found that most student´s reasoning rely on data and assimilate natural sampling variation, which are two fundamental ideas of inference. This result represents a significant change in their natural reasoning. An important misconception is believed Fathom simulates samples of the real population instead of a hypothetical one.

Keywords: High School Education, Technology, Informal Education.

Introduction

Literature shows many difficulties on the learning of statistical inference (Batanero, 2000) Castro-Sotos, Vanhoof, Van den Noorgate, & Onghena, 2007). One possible reason is that statistics courses generally focus on teaching procedures and routine concepts and do not offer the opportunity to discuss and understand the fundamental ideas. As a consequence, students reach their first inferential reasoning experience thinking that statistics is only about the computation of numerical values. This has motivated the interest in studying Informal Statistical Inference (IEI) and Informal inferential Reasoning (IIR). Researchers have recently been exploring the idea that if students begin to develop the informal ideas of inference early in a course, they may be better able to learn and reason about formal methods of statistical inference. In this context, the simulation (as opposed to the formal calculation) can be used to begin to teach the process logic and concepts that still need on the contrast of hypotheses (Batanero & Diaz, 2015). However, there are few studies on the RII aimed at students of high school (15-18 years).

Students present a lack of perception of sampling variation (García-Rios & Sánchez, 2014) and a lack of consideration of the data (García-Rios & Sánchez, 2015). This study seeks to show how a sampling distribution simulation activity has the potential to overcome these difficulties. In addition, this proposal presents a simulation by computer that can support the development of inferential reasoning for promoting the understanding of hypothesis tests. Therefore, we are interested in the questions: How students reason on significance testing when they participate in activities using the simulation of Fathom? How would Fathom support the learning of students?

Literature Review

Recently, several studies have focused on the concept of Informal Inferential Reasoning (IIR), as confirmed by the publications of special issues; Statistics Research Journal (Pratt & Ainley, 2008), and Mathematical Thinking and Learning (Makar & Ben-Zvi, 2011). Literature shows two different approaches to study the RII: the first approach focuses on the nature of reasoning about inference given problems and statistical information, while the second approach focuses on the evaluation of the development of the RII as students undergo a course of instruction designed to develop the reasoning. In this paper we focus on the first approach.

Garcia-Rios and Sánchez (2014) show that students have a lack of consideration of data and a low probability language; students often draw inferences based on personal beliefs instead of data and conclusions do not show any degree of uncertainty. In addition, when students based on data,

Galindo, E., & Newton, J., (Eds.). (2017). *Proceedings of the 39th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education.* Indianapolis, IN: Hoosier Association of Mathematics Teacher Educators.

they have an inappropriate way to determine if a statistic sample is significant; if the statistic is different from the hypothesis tested then it rejects the hypothesis. A plausible cause of these difficulties is the lack of perception of sampling variation (Garcia-Rios and Sánchez, 2015). These authors also observed that Fisher's test of significance comes more natural to students because they establish a null hypothesis (personal model of the population) to compare the sample and intuitively measure their significance (although inappropriate). It is concluded that in order to develop appropriate inferences before formalizing, it is crucial for students to have an informal method to determine a numerical criterion to know when rejecting or accepting the hypothesis and the simulation seems to be a resource that provides such method.

Rossman (2008) provides a characterization of informal statistical inference and makes a distinction between informal and intuitive, although it does not define it, just exemplified it by establishing some essential features of situations and problems of statistical inference and shows how it can informal methods be used to solve them. Zeiffler, Garfield, delMas, and Reading (2008) proposed a definition of the IIR and exposed three types of activities that should be generated by the tasks to develop it; they also propose a conceptual framework to characterize the RII and develop tasks that allow you to examine the natural IIR of students, as well as the development of such reasoning.

Conceptual Framework

In this work, the conceptual framework is understood as a network of related concepts or categories that together provide a general understanding of the phenomenon of research (Miles & Huberman, 1994).

Significance Tests

There are two different points of view about hypothesis testing: a) significance tests introduced by Fisher and b) testing rules to decide between two hypotheses, which was the opinion of Neyman and Pearson (Batanero, 2000). The approximation of Fisher emphasizes the strength of the evidence provided by the data observed against a null hypothesis. The strength of evidence is captured in the p-value, which measures the likelihood of having obtained an extreme result (or more extreme) if the null hypothesis were true. Under this assumption, the sampling distribution is calculated and from this distribution p-value is estimated; if the retrieved value is very small (statistically significant) the hypothesis is rejected.

Informal Inferential Reasoning

Several papers published in the last few years refer to the concepts of ISI and IIR, however there still no consensus as to what exactly these two concepts mean. In an attempt to combine the different perspectives, Zieffler, et al. (2008) defined the IIR as the way in which students use their informal knowledge of statistics to create arguments based on observed samples that support inferences about a population unknown. To emphasize the importance of informal reasoning we remember the ideas set by Bruner in 1960 (see Heitele, 1975) who believes that it is preferable that student begin to study the subject gradually, although initially only understand it either limited or informal, rather than wait until it matures and can teach directly in more abstract or formal. Teaching is not different in a structural way in the various educational stages, but only of a linguistic form and their level of deepening.

Method

This study is part of a Hypothetical Learning Trajectory (HLT) to develop students reasoning at the high school level. In this proposal, we focus on the reasoning of the students on significance

Galindo, E., & Newton, J., (Eds.). (2017). *Proceedings of the 39th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education.* Indianapolis, IN: Hoosier Association of Mathematics Teacher Educators.

testing with the use of technology while they completed a first task of a series of four, without having studied formal methods of statistical inference.

Participants

Thirty-six 11th grade (16-17 years of age) students, grouped in 18 pairs (referred to as R1 to R18) with a computer per couple, participated in the study. The participants had not studied statistics hence they lacked basic knowledge of statistics and never worked with Fathom. This means that the activities carried out have the objective of emerging student's insights on basic knowledge about significance testing and the use of Fathom.

Instruments

Data collection was conducted through a questionnaire applied in computer, in a two-hour class session. The data set are the answers given by pairs of students in a report about the conclusion of a proportion testing hypothesis problem. This report was written on the computer. The problem says "Coca cola claims that majority (more than 50%) of the population drinking cola prefer Coca rather than Pepsi. To check, an experiment where one gave two glasses of soda (one with Coca and other with Pepsi) to 60 people selected randomly from the population was done and they should decide what liked most. The 60 participants 35 people preferred Coke. Is the Hypothesis 'over the 50% of the population who drink cola in Mexico drink prefers that Pepsi Coca" correct?' Make a report were you: a) explain what your conclusion is: b) details how you came to your conclusion step by step: c) say what so sure of your conclusion are".

Process

Fathom's simulation and the problem were presented to students during the first hour to introduce and operate the software; generate random samples and sampling distributions. In the second hour, pair students were allowed to work freely to make a report of its findings (answers) on the computer; the teacher intervened only to answer small personal questions. When reports were finished students can leave class. Fathom simulates 500 samples (size 60) taken from a hypothetical population, where the parameter can be modified by a slider. Samples are represented in a bar graph and in a table (Figure 1). The simulation is used to generate an empirical sampling distribution and measure the likelihood of the observed data with the empirical method (frequency), i.e. the informal calculation of a p-value using frequencies (Rossman, 2008). The sampling distribution is shown in a table and a graph of points.

Results

Principles and techniques of *Grounded Theory* (Birks & Mills, 2011) were used to categorize students responses. This methodology claims that it is possible to develop emerging categories of data collected and analyzed systematically. The constant comparison of the data favors a full development of the categories and their properties (advanced coding) making it analytically powerful and therefore with the capacity to explain the phenomena under study. The categories of analysis that emerged from the data were: *sample, majority in simulation, mode in simulation and proto-significance test*. Each category reflects the different types of inferential reasoning posted in the student's reports. For analysis, responses were coded with the letter R and a number (table 1).

Galindo, E., & Newton, J., (Eds.). (2017). *Proceedings of the 39th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education.* Indianapolis, IN: Hoosier Association of Mathematics Teacher Educators.

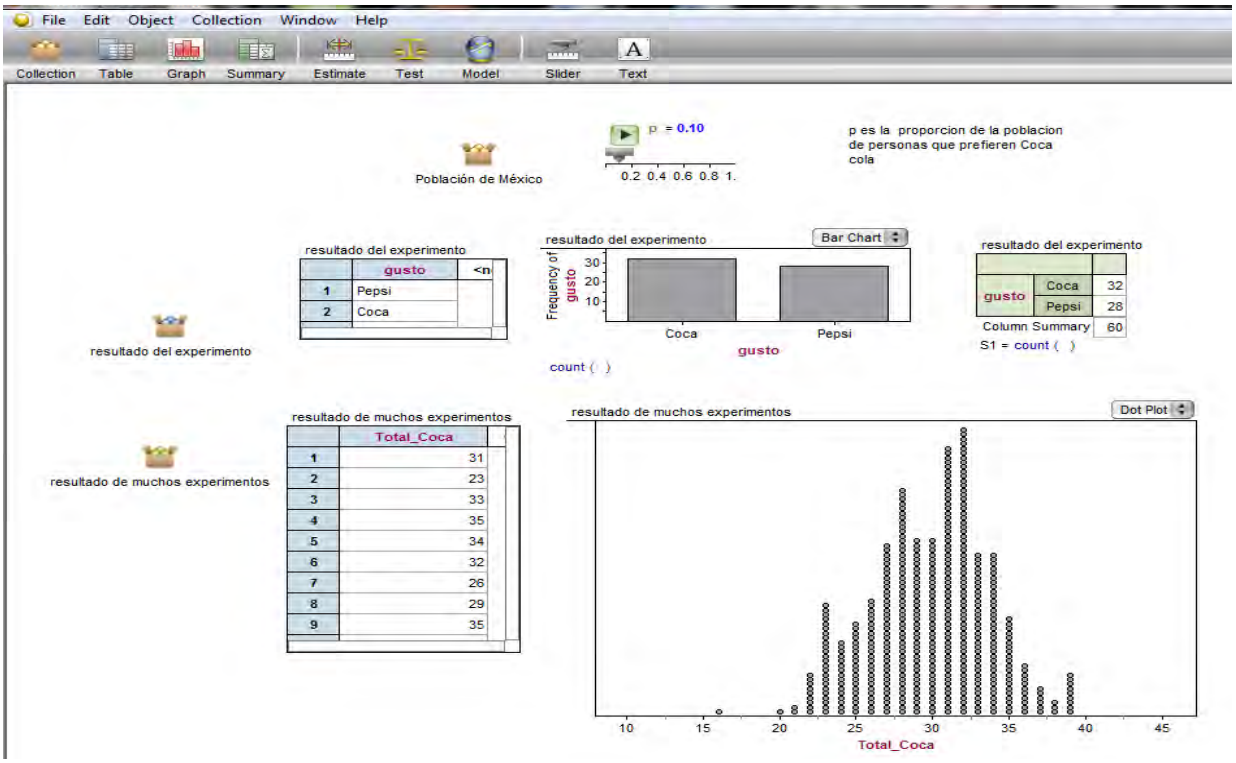

Figure 1. Fathom simulation screen.

The p-value of the statistic (0.58) is 0,098, so the hypothesis P = 0.5 is not rejected at a significance level of 5%. The categories of analysis that emerged from the constant comparison of data and explain the IIR students are: Sample, majority in simulation, mode in simulation and Proto-significance test.

Table 1: Students Reasoning

Category	Pair	Reasoning
Sample (10%)	R1, R3	If sample is bigger than 30 then hypothesis is incorrect
Majority in simulation (69%)	R2, R4, R5, R6, R7, R8, R9, R10, R11, R14, R16, R17	If majority of samples are bigger than 30 then hypothesis is correct
Mode in simulation (16%)	R18, R13, R15	If mode is bigger than 30 then hypothesis is correct
Proto-significance test (5%)	R12	Sample can occur within a population smaller than 0.5 therefore hypothesis can't be prove

Galindo, E., & Newton, J., (Eds.). (2017). *Proceedings of the 39th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education.* Indianapolis, IN: Hoosier Association of Mathematics Teacher Educators.

Sample

Responses within this category are solely based on sample data and do not consider the simulation results. This reasoning implies an absence of the idea of sampling variation. Therefore, if the proportion of the sample is greater (or smaller) 50%, students conclude that the hypothesis "more than 50% (less than 50%)" is correct. An example of this type of response is the pair R3 who explains in his report: "... from 51% of the population we can say that it is the majority and because the result of the experiment showed that 35 people 60 prefer Coca Cola which is equal to 59% of the total (59% x 60) = 35.4) we can conclude that they are not wrong in what they claim since they are right". In its report, R3 added figure 2.

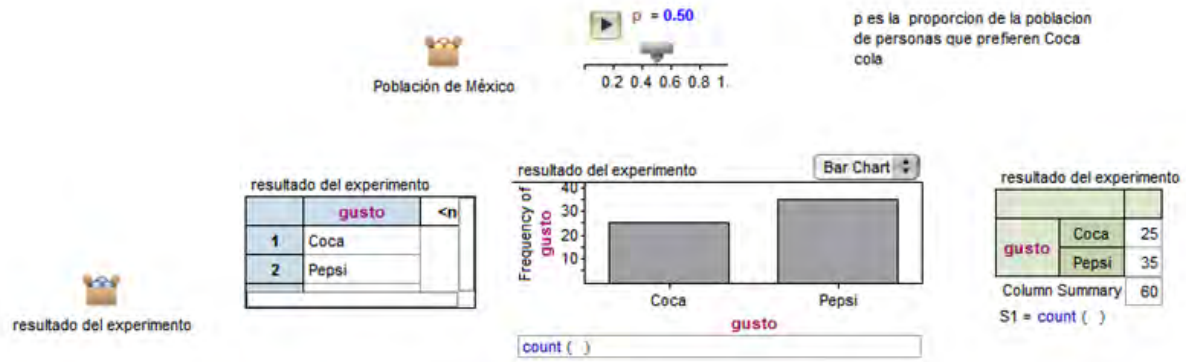

Figure 2. R3 Report.

Majority in Simulation

In this category the reasoning is to divide the sampling distribution in two regions; samples greater or equal to 30 (50%) represent the region that supports the hypothesis "most prefer Coca Cola" as correct (evidence against the null hypothesis), while less than 30 results represent the region which does not support the hypothesis. Thus, students come to their conclusion determining in which of these two regions are the most of samples. For example, R9 reason "... because if you select the rank of 37 we have 262 surveys... ". These couples add figure 3 and come to the conclusion that the hypothesis is correct. This reasoning suggests that students think that the results of the simulation with Fathom represent the actual population rather than a hypothetical, in addition, assimilate the sampling variation why they resort to determine in which region the most samples are. Some students use parameters greater than 0.5; R11 and R16 used P = 0.51, R4 use P = 0.54, and R10 and R14 use P = 0.6.

Mode in Simulation

The reasoning in this category focused on the mode of the simulated sampling distribution; If mode was greater than 30 (50%) then considered the hypothesis as correct (evidence against the null hypothesis). An example is R13 whom considered that the hypothesis is false, and reason "... our highest value in a survey was 29 people of 60 that liked more Coca-Cola, then from this we see that within that sample less than 50% like Coca-Cola". R13 added Figure 4.

Galindo, E., & Newton, J., (Eds.). (2017). *Proceedings of the 39th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education.* Indianapolis, IN: Hoosier Association of Mathematics Teacher Educators.

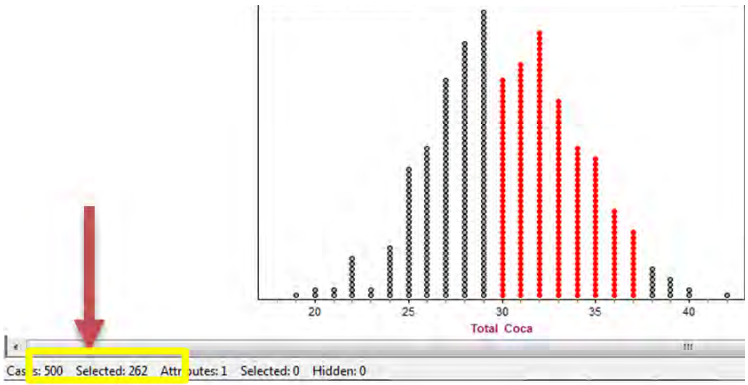

Figure 3. R9 report.

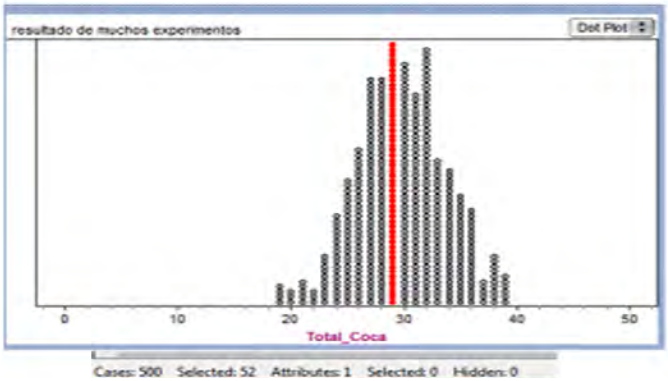

Figure 4. R13 report.

Proto-Significance Test

One interesting answer is the given by R12, whom conclude that it is not possible to test the hypothesis; "we must take a greater percentage of the population in general to the survey so we can conclude that more than 50% of the population actually likes or prefers Coca - Cola, because surveys within a range of 10 values greater and lesser around the expected value must be set" and continue "in the previous survey while the percentage is less than 50% (44%) a value of 26 is expected, I get results up to 16 (being the lowest) and 38 (being the highest) here we see a higher value that that the problem presents, where the majority of the population do not prefer Coca-Cola, so 35 does not ensure that the majority of the population like more Coca Cola". In other words, in a population less than 50% it is possible to obtain the sample; therefore the sample is not sufficient evidence to consider the hypothesis as correct.

Discussion

One of the principles of the constructivist approach applied to teaching is that for any new learning design the knowledge that the student already possesses should be used and articulated. Consequently, if it is to develop the students reasoning, is convenient to have the tools to know what knowledge and reasoning has and what are the false conceptions that limit them or blocked them. The answers to the research questions will give us knowledge to this end.

How students reason on significance testing when they participate in activities using the simulation of Fathom? The first important result is that all the arguments of the students were based on data. This is, no student based his reasoning on personal beliefs, difficulty found in (Garcia-Rios & Sánchez, 2014). The second result to highlight is the assimilation of the idea of sampling variation

by the majority of pairs of students (90%). It is considered that reasoning based on simulated sampling distribution have assimilated the sampling variation in some degree, when considering regions (variation of results of samples) and decide to take a statistician; majority or mode. An example of sampling variation assimilation explicit is R14 who wrote "although 50% of the population who likes Coca-Cola has been chosen, there are surveys were Pepsi wins the results". However, this assimilation is not sufficient to choose the outcome of the sample that rejects the null hypothesis (critical value). For a 5% level of significance should be 36, while students consider 30 (50%) or 31 (52%). This difficulty was also found in (Garcia-Rios & Sánchez, 2015).

The difficulties observed in the study are: 1) Sample-based reasoning. (2) A lack of variation to estimate the region that supports the hypothesis. (3) Although students use the simulated sampling distribution they didn't understand their role; responses suggest that students think that fathom simulation represents the actual population rather than a hypothetical.

How would Fathom support the learning of students? The use of absolute values and simulation of surveys made more visible and understandable abstractions such as the sampling distribution and its process. In the traditional teaching of the hypothesis testing a transformation of the statistics (typing or standardization) and the central limit theorem is used to calculate the p-value and determine the critic zone (of rejection) with help of the normal distribution. This is perhaps one of the darker aspects of all the techniques to students. The possibility of putting the notion of sampling distribution in the center of the discussion of a significance test is probably the main contribution of the use of educational software (in this case Fathom). In addition, the students described and explained the observed behavior instead of relying exclusively on theoretical arguments of probability, which often is counterintuitive for students (delMas, Garfield, & Chance 1999). These results show a path to follow for the development of the reasoning in the significance test. First, it must be understood that Fathom not simulates the actual population but a hypothetical; this can help to pass to the next level of reasoning. Secondly, based on the fact that many students assimilated sampling variation, discuss how to choose a sample result (critical value) to reject the hypothesis to verify, this will lead to the idea of p-value.

References

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

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.

Batanero C. & Díaz C. (2015). Aproximación informal al contraste de hipótesis. En 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.

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

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve student's statistical reasoning. *Journal of Statistics Education*, 7(3).

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

García-Rios, V. N., & Sánchez E. (2015). Dificultades en el razonamiento inferencial intuitivo. En 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.

Heitele, D. (1975). An epistemological view on fundamental stochastic ideas. *Educational Studies in Mathematics*, 6, 187-205.

Makar, K., & Ben-Zvi, D. (2011). The role of context in developing reasoning about informal statistical inference. *Mathematical Thinking and Learning*, 13(1–2), 1–4.

Miles, M. & Huberman, A. (1994). *An expanded sourcebook qualitative data analysis* (2aed.). Londres: Sage Publications.

Pratt, D., & Ainley, J. (2008). Introducing the special issue on informal inference. *Statistics Education Research Journal*, 7(2), 3–4.

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

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.