HIGH SCHOOL STUDENTS' MISCONCEPTIONS ABOUT SIGNIFICANCE TESTING WITH A REPEATED SAMPLING APPROACH

DIFICULTADES DE ESTUDIANTES DE BACHILLERATO SOBRE PRUEBAS DE SIGNIFICACIÓN A TRAVÉS DE UN ENFOQUE DE MUESTREO REPETIDO

Ernesto Sánchez Sánchez
Centro de Investigación y de Estudios Avanzados del IPN
esanchez@cinvestav.mx

Víctor N. García Rios
Centro de Estudios Tecnológicos en Aguas Continentales No. 09
nozairg@hotmail.com

Eleazar Silvestre Castro
Centro de Investigación y de Estudios Avanzados del IPN
eleazar.silvestre@gmail.com

Guadalupe Carrasco Licea
Colegio de Ciencias y Humanidades, Plantel Sur. UNAM.
gcarrascolic@gmail.com

In this paper, we address the following questions: What misconceptions do high school students exhibit in their first encounter with significance test problems through a repeated sampling approach? Which theory or framework could explain the presence and features of such patterns? With brief prior instruction on the use of Fathom software to generate empirical sampling distributions, 18 pairs of high school students participated in a series of lessons involving four significance test problems addressed by a repeated sampling approach. Based on the analysis of students' responses to the first problem, we identified four misconceptions about significance testing. A framework to explain the misconceptions is conjectured.

Keywords: Misconceptions, significance testing, empirical sampling distribution (ESD), verificationism, skepticism.

Sampling and inference are fundamental statistical ideas (Burril & Biehler, 2011). Statistical inference involves using a sample to make a claim about the population from which it has been drawn, quantifying the uncertainty in the process, then making decisions based on the information it provides. It arises from the necessity to evaluate experimental outcomes in any scientific domain. To fulfill this purpose, statistical inference relies on mathematical knowledge and thinking. This implies that the notions of sampling and inference are important components of statistical literacy and should be considered as candidates for compulsory topics in mathematics education curricula (Watson, 2006). However, the topic of statistical inference is usually not introduced until the university level, where it is compressed in an introductory course on statistics and probability. This is a concern, since teaching the wide scope of concepts that constitute statistical inference in just one semester often leads to poor understanding by students, who may only retain a set of terms and procedures.

A potential solution to this problem is to introduce the notion of statistical inference at different levels of education, with the degree of formality adapted to each level. Recent calls from the statistics education community to promote and advance the understanding of statistical inference (Pratt & Ainley, 2008) have spurred interest in exploring ways to introduce concepts related to statistical inference at the primary and high school levels. This research trend is consistent with the idea that Heitele, invoking Bruner, formulated 45 years ago: "…that any subject can be taught effectively in some intellectually honest form to any child at any stage of development" and that fundamental ideas "differ on the various cognitive levels, not in a structural way, but only by their linguistic form and their levels of elaboration" (1975, p. 187). Additionally, the rapid development of technological tools

In: Sacristán, A.I., Cortés-Zavala, J.C. & Ruiz-Arias, P.M. (Eds.). (2020). *Mathematics Education Across Cultures: Proceedings of the 42nd Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Mexico. Cinvestav / AMIUTEM / PME-NA. https:/doi.org/10.51272/pmena.42.2020

in education have allowed for new ways of representing both mathematical and statistical ideas that make these ideas more accessible at pre-university levels (Biehler, Ben-Zvi, Baker, Makar, 2013).

The notion of significance testing is a critical element in the understanding of statistical inference, and is an application of an important principle of scientific thinking: "one does not have evidence for a claim if nothing has been done to rule out ways the claim may be false" (Mayo, 2018, p. 5). However, the idea of validating the outcomes of scientific research by searching for ways to demonstrate the falseness of the claims is contrary to intuition: According to Fischbein (1987), people's natural inclination is to look for confirmatory evidence. Given the importance of and difficulties in understanding the concept of significance testing, therefore, we contend that this concept should be introduced at the high school level, when students are capable of reasoning in a formal way. By developing an understanding of some pertinent components of significance testing earlier in their mathematical education, we propose that students would be in a better position to tackle problems involving statistical inference at the university and professional levels. Accordingly, we formulated the following hypotheses: 1) It is possible to design and implement a series of lessons for high school students that involve reasoning with and about significance testing. 2) The use of technological resources is key to the implementation of the former. 3) The understanding gained about students' reasoning in response to such problems may point to a better teaching approach for significance testing. Using these hypotheses, we formulated our research question:

What misconceptions do high school students exhibit in their first encounter with significance test problems through a repeated sampling approach using computational simulation? What intuitions may explain these misconceptions?

We are motivated to investigate students' reasoning about significance testing by the fact that on the one hand, significance tests are extensively used in a wide range of scientific research domains to evaluate the results of experimental outcomes (Haig 2016; Winch & Campbell, 1970), while on the other hand, many students and researchers have a tendency to misunderstand the objective of significance tests and to misuse the associated concepts and results (Cohen, 1994; Goodman, 2008; Morrison & Henkel, 1997). This tension has given rise to two contrasting reactions to the use of significance tests in experimental research: a) The development and refinement of concepts embedded in the test, as well as the ways they are used in experimental procedures (Mayo, 2018); b) a strong opposition to the use of significance tests and calling for their retirement from experimental research practices (Amrhein, Greenland, & McShane, 2019). This situation points to the need for a better understanding of the ways in which students reason about significance testing.

Background

One of the considerable challenges in the teaching of statistics at the college level is enabling students to rationally interpret the system of ideas that constitute significance testing (Castro-Sotos et al., 2009; Vallecillos, 1996). Batanero (2000) identified three main concepts involved in significance testing that students often misunderstand: 1) the nature of the test, 2) the nature of the *p*-value, and 3) the significance level. In their own study, Castro-Sotos et al. (2009) described two frequent misconceptions about significance testing: 1) it is viewed as a mathematical proof that establishes the truth of one of the two hypotheses, and 2) it is viewed as a probabilistic proof by contradiction: that is, if the null hypothesis is rejected, the *p*-value is the probability of making the wrong decision. Castro-Sotos et al. (2009) also detailed several misconceptions about the concepts of the *p*-value and significance level, including: a) the *p*-value is the inverse of the normative definition (the *p*-value represents the probability of the hypothesis being true, given the outcome of the sample); b) the *p*-value is the probability of a simple event (the *p*-value represents the probability of obtaining the observed outcome); and c) the *p*-value is the probability of the null or alternative hypothesis being true.

Lane-Getaz (2017) explored how early education may decrease the incorrect use (and possible abuse) of significance testing. The researcher examined learning outcomes related to statistical inference for social science students in an introductory course on statistics that incorporated randomization and simulation tasks. The study demonstrated progress made by the students during the course and highlighted several difficulties and misconceptions that could be tackled through instruction.

Other than the previous study, however, the following have not received much attention in the research on students' understandings and misunderstandings about significance testing: 1) students' reasoning about significance testing in the context of a classroom intervention, and 2) the use of digital resources to generate empirical sampling distributions (ESDs) as a support for learning about significance testing. Both are addressed in our current experiment.

Conceptual framework

Reasoning is any process whose objective is to determine the validity or plausibility of a proposition or result by means of certain premises (data or propositions). Mathematical reasoning, in a broad sense, can be elaborated at different levels of formality, given that premises can range from intuitions to firmly instituted axioms, and derivation processes can range from persuasive argumentations (examples, induction, analogies) to well-established mathematical and logical procedures.

Reasoning is accompanied by *sense-making*, which involves linking a new proposition or unexpected result to previous knowledge or beliefs held by the learner (Shaughnessy et al., 2009). Sense-making is supported by intuition which, according to Fischbein, "…expresses the fundamental need of human beings to avoid uncertainty" (1987, p. 28). Doubtful information and uncertain propositions prevent actions and reasoning. Intuitions are necessary to act and reason because they are considered to be true by the subject. While intuitions derived from perceptions of reality are generally correct, "mental representations, hypothetical ideas and solutions may be biased, distorted, incomplete, vague or totally wrong. To believe, however, at least temporarily, in such mental productions, a certain excess of confidence is required" (p. 28). In summary, in the reasoning process, people need to start from certainties, which is why they tend to trust their intuitions (beliefs or conceptions) more than what would be justified under objective evaluation. According to Fischbein, the excess of confidence in intuitions is a mechanism that allows reasoning to be carried out, but at the expense of the risk of spurious conclusions. Fischbein goes on to explain that people's cognitive mechanism for acting and reasoning in certain environments consists of producing a coherent structure (Gestalt) while preserving "facts and segments which fit together and to discard those which may disturb the unity" (p. 35). In particular, drawing from the psychology literature, Fischbein explains that people manifest a bias to confirmation and are reluctant to seek non-confirmatory evidence; when it does appear, they tend to ignore it. They also tend to not consider other plausible scenarios nor that the very same evidence can also account for alternative hypothesis. We can add that hypothetical deductive reasoning is a formal expression of the need to start from (assumed) certainties in a scientific and controlled manner. However, such reasoning requires maturity and practice, given that it does not usually appear in a spontaneous way in specific situations, even when subjects have reached the formal operational stage of development proposed by Piaget (Schmid-Kitsikis, 1983).

The logic of significance testing requires overcoming the cognitive tendencies described by Fischbein. Indeed, a null hypothesis is a hypothesis that one tries to reject, rather than confirm, as is often believed: for Fisher (…), the null hypothesis is "the hypothesis that the phenomenon to be demonstrated is in fact absent." An experimental outcome is deemed significant when the null hypothesis is rejected, and not significant when the null hypothesis is not rejected. A statistic is a function that assigns a number to each sample of a given size. In the simplest case (the one used in

this study), this may be the proportion of an attribute in the sample. Significance testing is made possible by the ability to model the sampling distribution of the statistic under the assumption that the null hypothesis is true. It should be noted that in order to understand the role of the sampling distribution, one must apply hypothetical deductive reasoning. Once a sample has been drawn from the population and a statistic is calculated, a probability of the given event occurring or a more extreme value is computed, under the assumption that the null hypothesis is true. This probability is known as *p*-value. If the *p*-value is low (generally <5%), the null hypothesis is rejected, and the result is deemed significant.

A repeated sampling technique using software and random simulations allows students to construct an approximation to the sampling distribution of the statistic, which is usually referred to as an empirical sampling distribution (ESD). To simplify presentation, we define the following terminology as related to ESDs: assume that N samples of size k are simulated with H_0:P=θ, where θ is the probability of success. In this instance, success means selecting an element from the population that presents the feature R, and θ represents the proportion of elements of the population with the feature R. We define the statistic X as the "number of success cases in the sample" and use the notation ESD(N,k,P=θ) to denote an ESD of that statistic.

Method

Participants and data. The following results arise from data corresponding to responses provided by a group of 36 high school students to the first of four activities in a series of lessons on significance testing. Students were arranged in 18 pairs throughout the intervention, each with access to a computer equipped with Fathom. At the time of the intervention, the students were enrolled in their second year of high school (16-17 years old) and had not previously taken a course on statistics or probability. Activities were selected from statistics textbooks and then modified to align with the participants' school level and the repeated sampling approach. While solving the problems, each pair wrote a report that detailed their analysis and solution to the task; these reports constitute our main data source.

Instruments. The first problem of the series of lessons appears below; Table 1 summarizes the statistical data and the solution to the task.

> Coca-Cola's advertising campaign claims that the majority of people (more than 50%) prefer Coca-Cola to Pepsi. To corroborate this, an experiment was conducted, where 60 randomly-selected participants tasted both beverages in a blind test. Thirty-five of the participants preferred the Coca-Cola. Based on these results, would you accept the hypothesis that more than 50% of people prefer Coca-Cola to Pepsi?

Table 1. Data and solution to the first problem

Null Hypothesis (H_0)	N	Rejection freq.	*p–value*	Conclusion
$P = 0.50$	60	$X \geq 37$	$p = 0.126$	H_0 is not rejected

The intervention began with an introductory session aimed to enable students to use Fathom to generate ESDs. This was followed by a series of four lessons, each of which featured a problem involving significance testing. From the second lesson on, the lessons unfolded as follows: First, the instructor facilitated a discussion about students' solutions to the previous problem, inviting students to explain the motives and rationales behind their procedures; second, students were tasked with a new problem that gave them the opportunity to apply newly-gained skills and understandings about significance testing. Students worked on the problems in self-selected pairs and described their solutions on a worksheet. During this time, the instructor's role mainly involved posing questions to stimulate inquiry and helping students overcome technical difficulties (not to provide normatively correct answers).

We rely on principles of the grounded theory methodology (Birks & Mills, 2011; Glaser & Strauss, 1967/2008) for coding students' answers to the problems. This is a general research methodology in the social sciences (Holton, 2008) that involves constructing theories or frameworks through the gathering and analysis of data (as opposed to analyzing the data using an existing theoretical framework). As previously stated, the data analyzed in this study were students' responses to significance test problems. These responses were digitally transcribed, then analyzed in order to generate a set of codes and categories. Several misconceptions were identified using this procedure.

Results

Students' responses were analyzed and coded in three general categories: reasoning, hypothesis, and conclusion. In this report, we will only examine the four misconceptions identified in the first category (reasoning) exhibited in responses to the first problem. A more thorough analysis of the data can be found in García (2017); however, we consider the following evidence sufficient in exemplifying the identified misconceptions.

It should be noted that the descriptions of the emergent patterns use the authors' language and not necessarily the terminology students actually used when describing these ideas. Each identified misconception is a result of the constant comparative method, which allows for abstraction of the subtle differences in students' answers.

Misconception 1: ***Majority in the ESD***. Students generate an ESD (500, 60, P=0.50, or similar) and identify the number of samples in which the statistic X takes a value greater than 30. If this number is greater than 250 (N/2), the null hypothesis is rejected, and if it is lower than 250 (N/2), then it is not. It should be noted that this procedure ignores information given by the sample and the significance level. This misconception is consistent with the belief that the ESD represents the population, or an approximation of the population. This misconception was the most frequent in the data, identified in 13 out of 18 responses, and can be exemplified by R6's answer:

> …we considered surveys in which we had at least 31 favorable cases for Coca-Cola, which was a total of 233 surveys… while surveys with 30 favorable cases, that is, 50% or less, represented 267 surveys… a quantity greater than 50%.

The pair concluded that it is not true that there is a greater preference to Coca-Cola over Pepsi and included Figure 1a in their response.

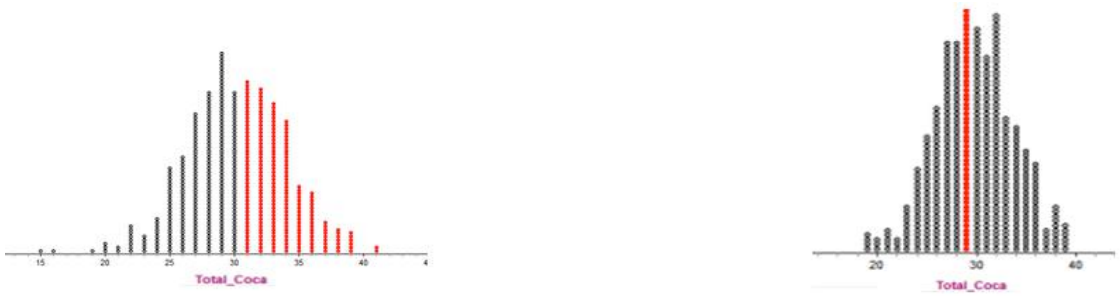

Figure 1: a) Majority in the ESD (pair R6); b) Mode (pair R13)

Misconception 2: ***Mode*.** Students generate an ESD (500, 60, P=0.50) and identify its mode. If this value is greater than 30, the null hypothesis is rejected, and if it is equal to or lower than 30, it is not. As in the previous case, this reasoning ignores both the observed outcome and the significance level, and is also compatible with the belief that the ESD provides some information about the probability of the null hypothesis being true: In other words, it is assumed that $p = frequency(H)$, where H is an independent variable. This misconception emerged in three responses. An example is provided by R13:

> …our most frequent value in the surveys was 28 people (out of 60) that liked Coca-Cola more, so we can see there are less than 50% that liked Coca-Cola the most.

The pair concluded that Coca-Cola's advertised claim is false and supported their argument with Figure 1b).

Misconception 3: ***Majority in the sample***. This is the only misconception in which students completely ignore the ESD and rely exclusively on the observed outcome (35/60 that preferred Coca-Cola). According to this line of reasoning, if the proportion of favorable to unfavorable outcomes is greater than 0.5, the null hypothesis is rejected, otherwise it is not. R2's response exhibits this misconception:

> …we can say that a majority is 51% or more, and the experiment's result showed that 35 out of 60 people preferred Coca-Cola, which is 59% of the total [(59% x 60) = 35.4]. Thus, we can conclude that they are not wrong in their conclusion.

The pair concluded that Coca-Cola's advertised claim is true, but unlike other students who participated in previous explorations of the task (Garcia & Sanchez, 2014), these students were unable to apply the ESD results generated with the software to support their argument.

Misconception 4. ***Extreme hypothesis*****.** In this case, $P > 0.5$ is taken to be the alternative hypothesis. Students arbitrarily establish a margin of error, M (e.g., 5% or 10%). Then, they pick a value for P such that $0.5 - M < P < 0.5$ and generate an ESD (500, 60, $P \neq 0.50$) to observe the values within the range of this distribution. If the value of the statistic at hand is in this range (which is the case for this problem), then the hypothesis $P > 0.5$ gets rejected. R10's response demonstrates this line of reasoning:

> …in general, we must take a greater number of surveys of the population so we can conclude that indeed more than 50% like or prefer Coca-Cola, because in these surveys there must be a range of around 10 values more and 10 values less around the expected value […] in the simulated survey, despite the percentage being lower than 50% (because P = 44%) and is expected to have a value of 26, we obtain results that go from 16 (the lowest value of X with a non-zero frequency) to 38 (the highest value of X with a non-zero frequency), from which we can see a greater value (for X) than in the original problem, which is why the 35 do not assure that most people like Coca-Cola.

As previously stated, these responses were provided to the first of the four problems in the series of lessons. It should be noted that the process of addressing students' ideas and solutions may have allowed them to gradually incorporate some previously absent but significant elements involved in significance testing, such as using all of the appropriate facts in the problem and determining if the observed outcome could be labeled as an outlier or not. Nevertheless, students were ultimately unable to provide solutions to this problem that were consistent with the logic of significance testing.

Conclusions and discussion

We propose an explanation based on Fischbein's (1987) theory about intuition for the first three observed misconceptions. According to Fischbein, intuition is related to people's tendency to avoid uncertainty and to look for confirmatory evidence. This idea, coupled with the traditional focus on proof in mathematics teaching, led us to propose a category in the context of our experiment called *naïve verificationism*, which consists of the belief that the objective of a significance test is to demonstrate the veracity of the null hypothesis. *Critical verificationism*, on the other hand, consists of the belief that the objective of the test is to compute the probability of the null hypothesis being true (or false). However, as students continue to participate in discussions about the problems and analyze solution strategies, they develop a kind of skepticism, which involves recognizing that it isn't possible to conclusively verify any hypothesis based on the evidence provided by a sample.

According to Fischbein (1987), people try to produce coherent structural schemata in which they can integrate their intuitions (certainties), beliefs (theories), and observations (evidence). Such structural schemata underpin the reasoning process—justifying the solution of a problem or explaining a phenomenon. In this sense, the misconception of mistaking an ESD with the population (or its approximation; Garfield & Ben-Zvi, 2008) is compatible with naïve verificationism, given that if the sampling distribution represents the population, then analyzing the ESD alone would be enough to verify the validity of the hypothesis. We interpret Misconception 1 (Majority in the ESD) as a manifestation of students' efforts to align the ESD, a new concept for our participants, to their expectations of verifying the null hypothesis. In a similar way, another way to create a structural schema is to believe that the ESD is a distribution in which each value for the statistic is taken as a possible hypothesis, with the frequency of each hypothesis (statistic) allowing for the computation of the probability of the truth (or falsehood) for each one; such reasoning demonstrates critical verificationism. The mode of such a conceived distribution would be the most likely hypothesis (statistic), and this is consistent with Misconception 2 (Comparing the mode of the ESD with the null hypothesis).

In Misconceptions 1 and 2, students do not assign a specific role to the information provided by the sample, given that their conception of the ESD makes this task an unnecessary one. These results show that the intended purpose and meaning of an ESD is not clear to students, despite their participation in activities in which they produced and interacted with different ESD's using physical and computational simulations.

In Misconception 3 (Majority in the sample), students do not ignore sample results as in previous cases, but create a simpler structural schema: they assume that because the information provided by the sample is the only information available when making a decision, it alone contains the key to the solution of the problem. Therefore, it is assumed that the proportion of favorable to unfavorable outcomes in the sample closely mirrors the corresponding proportion of the population. In this misconception, the information provided by the ESD is completely ignored, and therefore demonstrates naive verificationism. Shaughnessy (1992, p. 478) referred to two false conceptions that are related to this misconception: "people inadequately believe that there's no variability in the real world" and that "people often have an unjustified overconfidence in small samples." Nickerson (2000, p. 254) and, in a similar way, Castro-Sotos et al. (2009) report that one of the most common beliefs about significance tests among researchers and students is that "by rejecting the null hypothesis, a theory that predicts the falseness of the null hypothesis is established." Such a claim is compatible with naïve verificationism, because it responds to a desire of establishing the falseness of an hypothesis. This particular misconception emerged only in the first problem and was abandoned as students gained experience in generating and analyzing ESDs.

Furthermore, in Misconception 4 (Extreme hypotheses), students incorporate the idea of a margin of error by defining an interval on one side of the null hypothesis, with the opposite side representing the alternative hypothesis. An extreme hypothesis is selected in this interval. If the distribution generated by the extreme hypothesis captures the value of the statistic, there is no reason to reject the null hypothesis. That is, if the value of the statistic (in this case, 35) can plausibly occur in a more extreme ESD ($P < 0.5$), then the event is also plausibly occurring for the null hypothesis ($P = 0.5$) as well, and the null hypothesis should therefore not be rejected. In using such reasoning, students exhibit intuitions that are partially aligned with the logic of significance testing: for example, the idea of "taking more extreme values" is used when defining a p-value, and considering "what would happen if the hypothesis was…" demonstrates hypothetical-deductive reasoning that is used in the establishment of the hypothesis. However, such intuitions are not appropriately applied. An important feature of this line of reasoning is that it relies on an arbitrary, but well-intended criterion:

the null hypothesis is rejected when the statistic is not contained in a certain range of the simulated ESD.

In summary, the misconceptions that students exhibited in their solutions to the first problem correspond to the intuitive idea that the objective of a significance test is to verify the null hypothesis, or to estimate the probability that it is true. It is likely that a tendency to verificationism also explains some of the misunderstandings and misuses of ESDs evidenced even by experimented learners. This is why it is critical in the teaching of statistics to enable students to develop alternative reasoning schemes—that is, schemes based in reasonable skepticism.

Acknowledgments

Project financed by Proyecto-Conacyt 254301 and Proyecto SEP-Cinvestav 188.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance [Comment]. *Nature*, *567*, 305–307.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning, 2*(1-2), 75-98.

Batanero, C. & Díaz, C. (2015). Aproximación informal al contraste de hipótesis [Informal approach to hypothesis test]. 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. 135-144). Granada.

Biehler, R., Ben-Zvi, D., Baker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clements et al. (Eds.), *Third International Handbook of Mathematics Education* (pp. 643-689). New York: Springer.

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

Burril, G. & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In C. Batanero, G. Burril, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education. A joint ICME/IASE study* (pp. 57-70). New York: Springer.

Castro-Sotos, A. E., Vanhoof, S., Noortgate, W. V., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? *Journal of Statistical Education, 17*(2) (Retrieved from: https://www.tandfonline.com/loi/ujse20).

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003.

Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics 4*, 49-70.

Evans, G. (1997). A note on p-values. *Teaching Statistics, 19*(1), 22-23.

Fischbein, E. (1987). *Intuition in science and mathematics: An educational approach.* Dordrecht, The Netherlands: D. Reidel Publishing Company.

Fisher, R. A. (1960). *The design of experiments*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1973). Statistical methods for research workers. New York: Hafner.

García, V. N. y 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 y T. Ortega (Eds.), *Investigación en Educación Matemática* XVIII (pp. 345-357). Salamanca: SEIEM.

García, V. N. (2017). Diseño de una trayectoria hipotética de aprendizaje para la introducción y desarrollo del razonamiento sobre el contraste de hipótesis en el nivel medio superior. Unpublished doctoral tesis. México: Departamento de Matemática Educativa, CINVESTAV-IPN.

Glaser, B. G. & Strauss, A. L. (1967/2008). *Discovery of grounded theory: Strategies for qualitative research.* New Brunswick, USA: Aldinee Transaction.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology, 45,* 135-140.

Haig, B. D. (2017). Tests of statistical significance made sound. *Educational and Psychological Measurement, 77*(3), 1-18.

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

Holton, J. A. (2008). Grounded theory as a general research methodology. *The Grounded Theory Review, 7*(2), 67-93.

Lee, H. S. (2018). Probability concepts needed for teaching a repeated sampling approach to inference. In C. Batanero and E. J. Chernoff (Eds.), *Teaching and Learning Stochastics* (pp. 89-101), [ICME-13 Monographs]. New York: Springer.

Lane-Getaz, S. (2017). Is the p-value really dead? Assessing inference learning outcomes for social science students in an introductory statistics course. *Statistics Education Research Journal*, *16*(1), 357-399.

Lehman, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics.* New York: Springer.

Mayo, D. G. (2018). Statistical inference as severe testing: How to get beyond the statistics war. Cambridge, UK: Cambridge University Press.

Morrison, D. E. & Henkel, R. E. (Eds.). (1970). *The significance tests controversy: A reader*. Chicago: Aldine.

Nickerson, R. S. (2000). Null hypothesis significance test: A review of an old and continuing controversy. *Psychological Methods, 3*(2), 241-301.

Popper, K. R. (1990). *La Lógica de la Investigación Científica.* México D. F.: Red Editorial Iberoamericana.

Pratt, D., Ainley, J. (2008). Introducing the special issue on informal inferential reasoning. *Statistical Education Research Journal*, *7*(2), 3-4.

Schmid-Kitsikis, E. (1977). The development of hypothetic-deductive thinking and educational environments. In M. McGurk (Ed.), *Ecological factors in human development*. Amsterdam: North Holland Pub. Co.

Shaughnessy, J. M. (1992). Research in probability and statistics. In D. A. Grouws (Ed.)*, Handbook of research on mathematics teaching and learning* (pp. 465-494). New York: Macmillan.

Tukey, J. (1986). What have statisticians been forgetting? In L.V. Jones (Ed.), *The collected work of John W. Tukey. Volume IV: Philosophy and principles of data analysis, 1965-1986* (pp. 587-599). Monterrey, CA: Wadsworth & Brooks.

Vallecillos, A. (1996). *Inferencia Estadística y Enseñanza: Un análisis didáctico del contraste de hipótesis.* [Colección MATHEMA]. Granada, España: Comares S. L.

Watson, J. M. (2006). *Statistical literacy at school.* Mahwah, NJ: Lawrence Erlbaum Associates.

Winch, R. F. & Campbell, D. T. (1970). Proof? No. Evidence? Yes. In D. E. Morrison & R. E. Henkel (Eds.), *The significance tests controversy: A reader*. Chicago: Aldine.