MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2024, 20(1), em2388
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/14119

An approach to inferential reasoning levels on the Chi-square statistic

Jesús Guadalupe Lugo-Armenta [1], Luis Roberto Pino-Fan [1*]

[1] Universidad de Los Lagos, Osorno, CHILE

Received 03 November 2023 ▪ Accepted 27 December 2023

Abstract

This paper presents an approach of progressive levels of inferential reasoning on the Chi-square statistic, going from informal to formal reasoning. The proposal is based on epistemic criteria retrieved from a historical-epistemological study of such statistic and the contributions of statistics education literature on inferential reasoning. In this regard, some theoretical and methodological notions from the onto-semiotic approach were used to identify meanings attributed to the Chi-square statistic throughout its evolution and development. The mathematical characteristics of those meanings are closely linked to the indicators of the levels proposed. The nature of the four levels on the Chi-square statistic allowed us to develop an initial approach to levels of inferential reasoning, which could be applied to other statistics such as z, student's t and F.

Keywords: Chi-square, inferential reasoning, statistical inference, statistics education

INTRODUCTION

In the last decades statistics has acquired greater relevance in the curricula and researchers have shown an increased interest in new ways of approaching statistical inference. In the scientific literature on statistical education, it has been proposed that the notions of inference could be introduced informally from an early age, seeking to familiarize students with these notions, so that when they approach topics of statistical inference they can understand integral formal notions, procedures and language.

Among the proposals to introduce inference informally, activities that promote the idea of prediction, generalization, the relationship between sample and population, mean, sampling distributions, sampling variability, distribution and probability stand out (Doerr et al., 2017; English & Watson, 2018; Makar, 2016; Pfannkuch, 2007; Rossman, 2008). In addition, research has emerged on an informal approach to hypothesis testing, confidence intervals, correlation and linear regression, and analysis of variance (e.g., Dierdorp et al., 2011; Dolor & Noll, 2015; Stohl Lee et al., 2010; Trumpower, 2013, 2015; Weinberg et al., 2010).

In the Chilean mathematics curriculum, as has happened in the curricula of various countries, the topics of statistical inference are now found at the pre-university level and students are expected to be able to make inferences about the mean and variance by applying procedures confidence intervals or hypothesis tests (Mineduc, 2019). In this context, on the one hand, research has been carried out in primary education on early learning of statistics and probability, these studies involve aspects of informal inferential reasoning (IIR), literacy and statistical reasoning (e.g., Vasquez & Alsina, 2019; Vidal-Szabó et al., 2020). On the other hand, studies have been carried out in secondary education on IIR (e.g., Estrella et al., 2023), elements of informal inference have been identified in eighth grade mathematics textbooks (Sanchez & Ruiz, 2022) and on inferential reasoning with mathematics teachers (e.g., Lugo-Armenta & Pino-Fan, 2021c, 2022).

Some proposals have developed on novel ways of addressing inference, both from the standpoint of informal inference and from a movement toward formal inference. On the one hand, the perspective of IIR aims at integrating and giving meaning to statistical notions, as well as developing a preliminary approach to Inferential Statistics (e.g., Doerr et al., 2017; Makar & Rubin, 2009; Zieffler et al., 2008). Regarding this, Zieffler et al. (2008), propose a framework focused on three types of tasks to develop IIR, as well as the ways of thinking involved in it, they include: estimating and graphing a population, comparing two samples of data, and judging

✉ jesus.lugo@ulagos.cl ✉ luis.pino@ulagos.cl **(*Correspondence)**

Contribution to the literature

- The proposal for progressive levels of inferential reasoning, from informal to formal, on the Chi-square statistic, based on the mathematical richness retrieved from the historical-epistemological study on this statistic and of the Statistics Education literature on inferential reasoning.
- The 'indicators' of the levels of inferential reasoning can serve as an initial guideline both for lesson planning, the design of activities that progressively promote inferential reasoning, and for studies on the level of inferential reasoning that is promoted in the mathematical practice of students.
- Another contribution of this research is related to the identification of a future line of research that allows characterizing the inferential reasoning.

between two competing models. Each of the three tasks incorporates the three components of IIR:

(1) making judgments or predictions,

(2) using or integrating prior knowledge, and

(3) articulating evidence-based arguments.

Simultaneously, Makar and Rubin's (2009) framework, which is based on concepts they initially saw as critical, propose three principles that are essential in informal inference:

(1) generalization, going beyond describing the data;

(2) using data as evidence in generalizations; and

(3) using probabilistic language in describing the generalization.

On the other hand, some research (e.g., Jacob & Doerr, 2014; Makar & Rubin, 2018; Pfannkuch et al., 2015), suggests that students should be introduced to inference in stages, in other words, that they should be taught to use formal inferential reasoning (FIR) in a progressive fashion. Pfannkuch et al. (2015) indicate some fundamental concepts, for informal inference ('making a call', sample-population ideas, sampling variability) and for formal inference (bootstrap method and randomization method), which can be worked at various moments in the school curriculum. They also present some learning activities comprising the comparison of boxplots and the use of software that help explain each of the fundamental concepts. However, the question of how we can build a FIR on the basis of IIR is still being debated.

Topics such as hypothesis testing, confidence intervals, and so on, are usually approached in a formal way in university courses and, according to several investigations (Batanero et al., 2012; Garfield & Ben-Zvi, 2008; Harradine et al., 2011; Makar & Rubin, 2018; Sotos et al., 2007), both, students and professors have presented difficulties when working on inference topics. According to Bakker and Derry (2011) and Makar and Ben-Zvi (2011), one of the main problems facing the teaching of statistics is that notions are taught in isolation (both from each other and from the context from which the data arise). For example, notions such as the Chi-square statistic, its distribution, probability, significance, p-value, among others, are commonly taught with a predominantly algorithmic approach (Batanero, 2013; Matis et al., 2004) and without emphasizing the connections between them; this contrasts with the holistic approach required for statistical reasoning. Thus, the importance of generating an initial approach to statistical inference is recognized (Bakker & Derry, 2011; English & Watson, 2018; Makar et al., 2011; Zieffler et al., 2008). Therefore, creating learning opportunities that facilitate the progressive development of the understanding of the key notions of statistical inference is critical.

In order to progressively promote a FIR, it is essential to understand how statistical notions emerge from the mathematical practices that have contributed to the solution of different types of problems, as this makes possible to identify the different meanings of the same notion. In this study, we take as an example the Chi-square statistic, for its importance in the application of inferential statistics and in statistics education, and which hypothesis tests are part of applied statistics; contributing significantly to medicine, psychology, genetics, aquaculture, biology, financial analysis, econometrics, industry and marketing research. Despite the importance of the Chi-square statistic, making inferences based on its tests requires a deep understanding of the statistic and the notions that are related to it, because although students may perform the procedures seemingly adequately, they may have difficulties in understanding or connecting the statistical notions involved. In this regard, several studies have documented errors and difficulties in the understanding of the Chi-square statistic (e.g., Cañadas et al., 2012; Vallecillos, 1994; Vera & Díaz, 2013; Vera et al., 2011). The findings include difficulties in interpreting in the context of the problem, confusion between the critical value and the p-value, recognizing the value of the probability associated with the value obtained from the statistic, confusion between the degrees of freedom and the homogeneity test with the independence test, the null and alternative hypotheses, confusion between the parameter and the statistic, etc. In addition to that, on a previous study, we have identified various meanings attributed to the Chi-square statistic throughout its evolution and historical development. These meanings provided epistemic criteria for the construction of levels of inferential reasoning on the same statistic.

Based on the previous, the objective of this article is to present a proposal of progressive levels of inferential reasoning, from the informal to the formal, on the Chi-square statistic, based on both the mathematical richness retrieved from the historical-epistemological study on this statistic, and the contributions of the literature of statistics education on inferential reasoning. The proposed IIR levels present adjustable *indicators* in generalization and formalization processes, which allow designing activities that progressively promote inferential reasoning, as well as performing detailed analyses of mathematical practices (curriculum, textbooks, students, etc.) to determine the levels of IR being promoted.

THEORETICAL-METHODOLOGICAL FRAMEWORK

For the development of this study, we have used the onto-semiotic approach (OSA) to mathematical knowledge and instruction (Godino et al., 2007, 2019). Presmeg (2014) points out that OSA is an inclusive theoretical system that tries to articulate various approaches and theoretical models, which allow studying phenomena in the research in mathematics education, referring to at least six large dimensions (in OSA called facets) present in the processes of teaching and learning mathematics: *epistemic facet* (relative to the mathematical richness and complexity of a given mathematical content, i.e., institutional meaning of reference, planned or implemented of the mathematical content); *ecological facet* (the connections of the content to be taught with other areas, as well as the social, political, economic factors, etc., that condition the teaching and learning processes); *cognitive facet* (regarding the level of development of the students, that is, comprehension and mathematical competence, difficulties and errors in the study of the content); *mediational facet* (material, technological and temporary resources, their uses to promote the teaching of content, their impact on the stability of the personal meanings that students achieve); *interactional facet* (organization of discourse in the classroom and of the interactions between the agents involved in the teaching and learning processes of the contents, considering the learning difficulties of the students and the negotiation of meaning); *affective facet* (Emotions, attitudes, beliefs, values, interests and needs of students regarding the mathematical content).

In this article, aspects related to the epistemic facet are of special interest, particularly in relation to the epistemological and ontological problem summarized with the questions How does mathematics emerge and develop? What is a mathematical object? What types of objects are involved in mathematical practices? In order to answer the epistemological problem, an anthropological (Wittgenstein, 1953) and pragmatist (Peirce, 1958) vision of mathematics is assumed; therefore, the activity of people in problem solving is considered the central element in the construction of mathematical knowledge (Godino, 2022; Godino et al., 2007; Pino-Fan et al., 2017).

The notion of *mathematical practice* takes on a fundamental role in OSA and is understood as "any performance or manifestation (verbal, graphic, etc.) carried out by someone in order to solve mathematical problems, to communicate the solution to others, to validate the solution and to generalize it to other contexts and problems" (Godino & Batanero, 1994, p. 334). The practices can be idiosyncratic of a person (personal practices) or shared within an institution (institutional practices), but indeed, it is the operative and discursive practices of people that, in solving certain types of problems, give rise to 'mathematical knowledge'. Mathematical practices involve ostensive objects (symbols, graphs, etc.) and non-ostensive objects (concepts, propositions, etc.), which we evoke when doing mathematics and which are represented in textual, oral, graphic or even gestural form. From the systems of operative and discursive mathematical practices, at least six new objects (or primary entities) emerge that come from them and account for their organization and structure: linguistic elements (representations), situations/problems, concepts/definitions, properties/propositions, procedures and arguments (Godino et al., 2007, 2019). These primary entities, which in OSA are called *primary mathematical objects*, interact to shape mathematical activity.

Particularly, *situations-problems* are the starting point or basis of the activity; *language* allows for the representation of the remaining entities and serves as instrument for action; *arguments* justify the procedures and propositions that connect concepts to one another. Godino et al. (2011), indicate that these primary mathematical objects can be analyzed from a process-product perspective, which implicates considering the following processes: communication (using linguistic elements), problematization (types of problems), definition (of concepts), enunciation (of propositions/properties), algorithmization (allows to elaborate the procedure), and argumentation. Other processes in OSA that allow to understand the complex and progressive nature of mathematical objects are generalization, particularization (exemplification), materialization, representation, modeling, idealization (outlining), signification, reification and decomposition (Font & Rubio, 2017; Medrano & Pino-Fan, 2016).

So, OSA recognizes a dual nature for mathematics: as a system of objects and as a system of practices. OSA adopts the anthropological and pragmatism view of socio-epistemic relativism of the system of practices, of the emergent objects and meanings of them, which allows addressing the semiotic-cognitive problem summarized with the questions: What does a mathematical object *O* mean for a person or institution at a given moment and context? What is it to know a mathematical object? Hence, when we ask, what is the Chi-square statistic? What is the student's t? And in general, what is the meaning of a particular mathematical object? OSA proposes that: it is the system of practices that a person carries out (personal meaning), or that is shared within an institution (institutional meaning) to solve a type of situations-problems. In this way, it is evident that a certain mathematical object does not constitute a single idea to be taught, but each mathematical object, each content, has multiple meanings, adjustable in processes of generalization and formalization, which emergence is progressive as a result of the activity (practice) of people at different times to respond to different situations-problems (Font et al., 2013). Each mathematical concept is understood in an anthropological, pragmatic and systemic way, implying various partial meanings (PM) or senses (e.g., Pino-Fan et al., 2011, 2017, 2018), which relate in a complex way and address the contexts in which they are used and institutional frameworks. Godino (2022), points out the following:

> As a basis for the didactic analysis, it is considered necessary to reconstruct a global or holistic meaning of the mathematical object through the systematic exploration of the contexts of use of the object and the systems of practices that are put into play for its solution. Said holistic meaning is used as an epistemological and cognitive reference model of the PM or senses that said object can adopt and constitutes a methodological tool for the onto-semiotic analysis of cognition: A method to delimit the various meanings of mathematical objects, and, Therefore, for the reconstruction of the epistemological and cognitive reference models, it is the analysis of the systems of practices (personal and institutional) and the onto-semiotic configurations involved in them (p. 9).

On that basis, we could say that in OSA the *reasoning* is assumed as a "social and epistemic macro-process", which involves putting into play both the primary mathematical objects, and the aforementioned processes, to solve a situation-problem (e.g., Lugo-Armenta & Pino-Fan, 2021b; Godino et al., 2015; Molina, 2019). Then, to say that a subject 'understands' the χ^2 we must observe that in his reasoning associated with his practices (to solve different types of situations/problems), primary mathematical objects and processes linked to the meanings of this notion, emerge gradually, systematically and progressively. In order to characterize inferential reasoning using the levels of inferential reasoning presented in this article, then, it is done in terms of the types of mathematical/statistical tasks, objects, and processes involved in the practice.

MEANINGS OF CHI-SQUARE STATISTIC

The historical-epistemological study on the Chi-square statistic has revealed major issues that were crucial to its genesis, development, and generalization. Distinct problem-situations, linguistic elements, concepts/definitions, properties/propositions, procedures, and arguments were mobilized in the mathematical practices used to address these difficulties, allowing different (partial) meanings to be identified.

Four main meanings for the Chi-square statistic (Lugo-Armenta et al., 2021) have been identified, which are made up of twelve PM (**Figure 1**). We will use the notations M1, M2, M3, and M4 to refer to the respective identified meanings; similarly, we will use PM1, PM2, PM3, and so on, to refer to the corresponding PM.

A goodness-of-fit test with the χ^2 statistic, as hypothesis testing, seeks to assess the extent to which a group of observed data is adjusted to a certain pre-established theoretical distribution, by means of the contrast of observed and expected frequencies. The graphic method (PM1χ^2) corresponds to an intuitive goodness-of-fit test, where the deviations from each sample data are analyzed from the mean value of the same sample. The second meaning, the test of independence, allows us to determine whether two variables are associated. It should be noted that in PM6χ^2 (test of independence through contingency with χ^2) the notion of association was generalized (from PM5χ^2–association test with Q coefficient–), giving rise to the test of independence with the χ^2 statistic. Although the test of independence and the test of homogeneity test are mathematically identical, the latter, determines whether it is possible that several samples come from the same population. As the χ^2 statistic evolved, to solve various problems, the χ^2 distribution also evolved, since Pearson (1900) obtained it as the asymptotic distribution of the χ^2 statistic when working on the goodness-of-fit problem for a frequency curve.

From this historical-epistemological study we identified primary mathematical objects, from the various PM, that could constitute possible 'paths or trajectories' for each major meaning of this statistic.

For example, if we consider the goodness-of-fit test (M1χ^2) of the Chi-square statistic, we note that this type of fit could be assessed initially with more intuitive methods (first stage of the path), such as the intercomparison method and the graphical method (Galton, 1875, 1885) of the PM1χ^2. From the PM1χ^2 we highlight primary mathematical objects such as the concepts/definitions ogive, quartiles, percentiles, deviations; properties/propositions normal distribution, m at $\frac{1}{2}$, p at $\frac{1}{4}$, q at $\frac{3}{4}$, probable error, and $q - m = m - p$. Following the path indicated by the historical study, in a second stage the goodness-of-fit could be assessed by interpreting the probability, obtained from the calculated Chi-square statistical value,

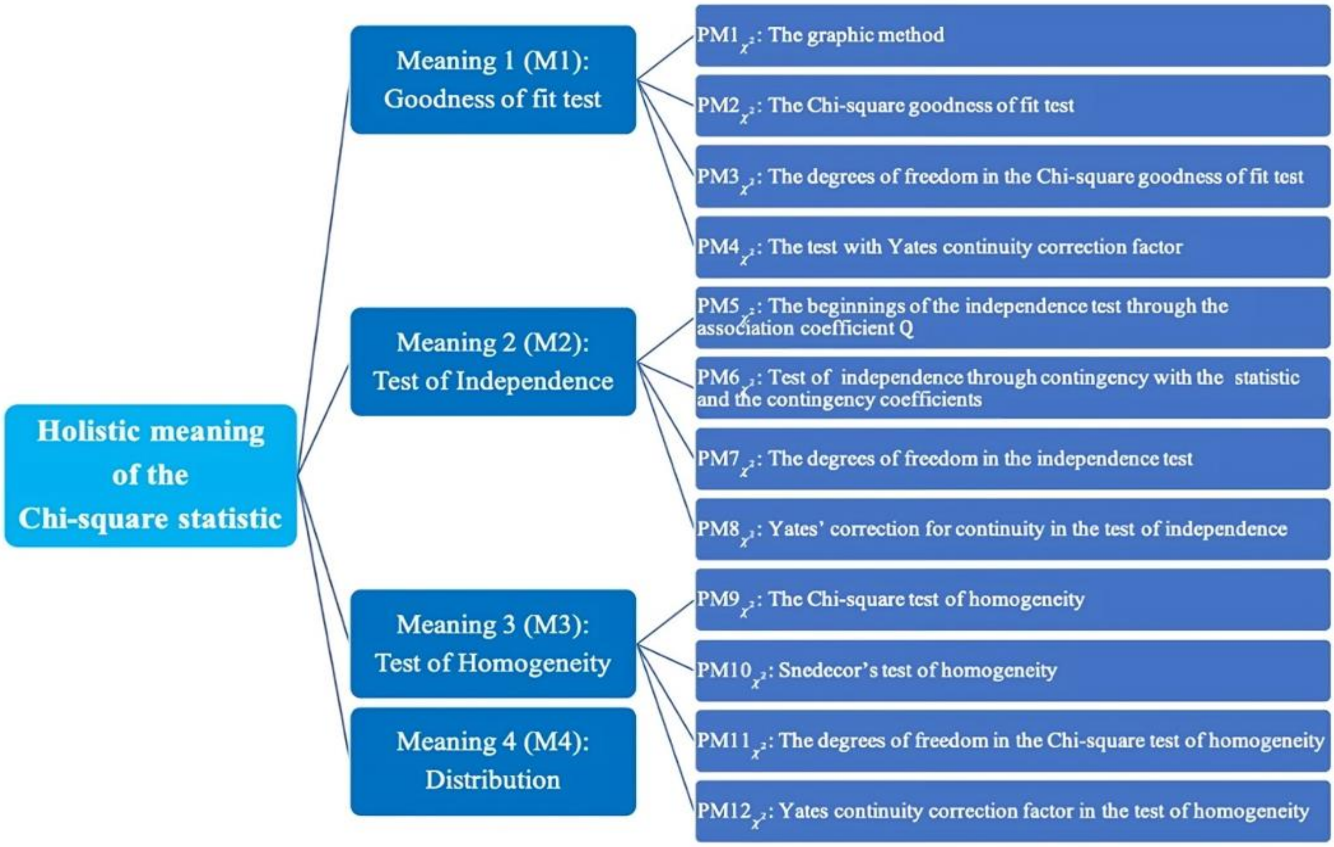


Figure 1. Meanings of the Chi-square statistic (adapted from Lugo-Armenta et al., 2021)

as a measure of occurrence of a complex system of n errors occurring with a frequency as large or larger than that of the observed system. In addition, the hypothesis implicit in the problem could be identified and stated in natural language. In this case primary mathematical objects of PM2χ^2 and PM3χ^2 are involved, of which the concepts/definitions of observed frequency, theoretical frequency, probability, and the properties/propositions $e = m' - m$, Chi-square probability distribution and Chi-square statistic of PM2χ^2 stand out. The same happens with the concepts/definitions degrees of freedom, significance and the properties/propositions $k = n - r$ and decision rule of PM3χ^2. In a third stage, we could introduce the limitation of using the Chi-square statistic when the sample we are analyzing has expected frequencies of less than five and the use of the Chi-square statistic with continuity factor. In addition, one could approach significance as indicative of the level at which the possibility of the effect should receive serious consideration and apply a decision rule to conclude on whether the observed data conform to the expected theoretical distribution. In order to carry out what was proposed in this third stage we highlight the relevance of primary mathematical objects such as the concept/definition continuity correction factor and the property/proposition $\chi_c^2 = \sum \frac{(|m'-m|-0.5)^2}{m}$ of the PM4χ^2 and those pointed out in the second stage. In the fourth stage of the trajectory, the goodness of fit could be assessed based on the statistical techniques of the hypothesis testing methodology, for which the statistical hypotheses could be formulated, using the Chi-square statistic, with the continuity correction factor if necessary, finding the probability value associated with the value of the statistic and using the decision rule (p-value or critical value) to conclude whether the observed frequencies are distributed as expected. At this stage of the trajectory, we highlight the primary mathematical objects previously described in the third stage and the decision rule (property/proposition of the PM4χ^2).

Each of the stages of the trajectories, previously exemplified for the goodness-of-fit test, constituted a significant input for the construction of our proposed levels for the Chi-square statistic. In the following section we present the primary mathematical objects that make up each of the trajectories linked to the indicators of the levels of inferential reasoning.

INFERENTIAL REASONING LEVELS FOR THE CHI-SQUARE STATISTIC (χ^2)

In this section, we present four levels of inferential reasoning for the χ^2 statistic, which combine the contributions of the statistics education literature, such as the primary mathematical objects, and the processes identified in each of PM emerging from the historical-epistemological study. Then, the essential characteristics (primary mathematical objects) of PM will be enunciated based on their gradualness in generalization -i.e., from

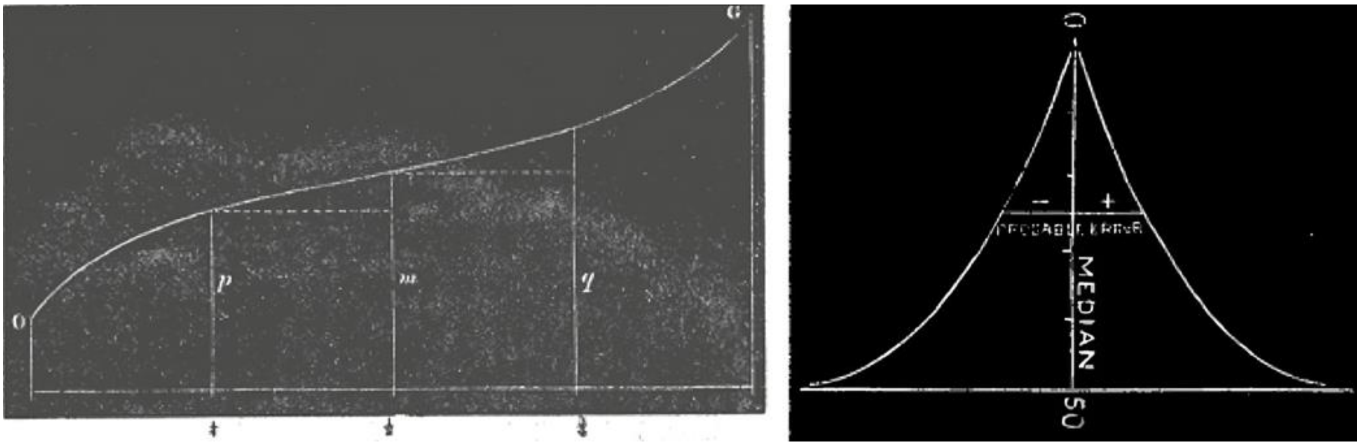


Figure 2. Graphics belonging to PM1χ^2 of goodness-of-fit-test (Galton, 1875, 1885)

the particular to the general- and formality -i.e., from the intuitive to the formal-, which gives an account for the progressivity of the proposed levels.

It is important to highlight that this four-level proposal considers a transition from level 1 (informal) to level 4 (formal), considering the transit through levels 3 and 4 (pre-formal). In this sense, there are indicators of different levels that are linked, for example, at levels 2, 3, and 4 indicators are presented on the approach of statistical hypotheses, ranging from the identification of the null hypothesis implicit in the problem and its approach in natural language, up to the approach of the null and alternative hypotheses with statistical language. Thus, when considering progressive indicators, the levels of inferential reasoning complement each other. However, given the nature of a certain educational level, only one level could be used, for example, in higher-level Statistics courses at the university level, the fourth level indicators can be used both for class planning and design, and to characterize the inferential reasoning that students exhibit in their mathematical/statistical practices. While in secondary education (14 to 18 years) the indicators of levels 1, 2, and 3 can be used.

Level 1–Informal

This level provides 'indicators or elements' that correspond to an IIR, where the visualization of graphs is used to establish conjectures and then, data is analyzed. Mathematical objects from descriptive statistics and probability intervene to make substantiated conjectures.

Visualization

The student is expected to be able to conjecture and argue whether the data of a sample follow a normal distribution through the elements present in the graphs (e.g., shape, dispersion, quartile amplitude, median, skewness). This, in a similar fashion to how Zieffler et al. (2008) tackled the first portion of task 1 to promote IIR. **Figure 2** shows the type of charts that can be presented to students or that they could graph from the observed frequencies.

Working with data

The student can analyze the data in a sample under Galton's graphical method (Galton, 1875, 1885) to conjecture whether the data group follows a normal distribution; this can be seen in two parts.

The first corresponds to the method called intercomparison, involving primary mathematical objects of PM1χ^2 (e.g., concepts/definitions such as ogive, quartiles, percentiles and deviations; and properties/propositions such as normal distribution, m at $\frac{1}{2}$ or 0°, p at $\frac{1}{4}$ or $-25°$, q at $\frac{3}{4}$ or 25°, and conditions for a symmetric series). In the second part of the graphical method, the student can use the results of the first part to calculate and graph the deviations, positive and negative, with respect to the median, for which he makes use of the mean error concept and the probable error property of PM1χ^2. It is possible to make the connection of deviations from the median with the error ($e = FO - FE$) of the χ^2 statistic of Pearson (1900) (PM1χ^2 and PM2χ^2).

Intuitive association

The student can establish whether there is an association between two variables, with an attribute of each variable, by means of the coefficient of association Q. Specifically, from this sublevel it is possible to observe in the practices the use of:

- Basic concepts/definitions of probability and their properties –e.g., variable, qualities or attributes, frequency, probability sets, association, $(AB)(U) = (A)(B)$, and $(AB)(\alpha\beta) = (A\beta)(\alpha B)$,
- Representation of data with 2×2 contingency tables is shown in **Table 1**.

Table 1. 2×2 contingency table

	(B)	*(β)*
(A)	*(AB)*	*(Aβ)*
(a)	*(aB)*	*(aβ)*

- The process of relating the property of the coefficient of association Q to Pearson's correlation coefficient concepts because the meaning of the association with this coefficient is in terms of correlation for continuous variables. Eq. (1) shows Q.

$$Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}. \quad (1)$$

The primary mathematical objects found at this level correspond to PM5χ^2 of M2χ^2.

It is important to remember the importance of working with types of situations-problems in contexts close to students, since various research (e.g., Bakker & Derry, 2011; Bakker et al., 2017; Ben-Zvi & Aridor-Berger, 2016; Gil & Ben-Zvi, 2011; Makar & Ben-Zvi, 2011; Makar et al., 2011) have highlighted the importance of context in promoting an IIR in students. This type of reasoning is not only based on the statistical knowledge of students, but also considers the role of informal reasoning and informal knowledge, as stated by Zieffler et al. (2008).

Level 2–Pre-Formal

Some 'indicators or elements' of this level have IIR features, such as how it approaches the null hypothesis, as it is implicit in the problem, and the meaning given to probability. That is why this level can be considered as pre-formal.

These ways of approaching the tests can be seen in history, for example, at the beginning of the goodness-of-fit test, Pearson (1900) asked a question about the problem he was raising; we can see this question as the null hypothesis.

Identify non-parametric test needed to analyze data

In order for a student to identify the appropriate χ^2 test to analyze the data, the following indicators need to be observed in the practices:

- Recognize the type of data you are working on (e.g., whether it is a sample or a population, are qualitative or quantitative, are classified according to one or two variables, sample type and number of samples).
- Understand the problem to be solved.
- Understand and enunciate the uses of the goodness-of-fit test (M1χ^2), independence (M2χ^2) and homogeneity (M3χ^2). For example, the goodness-of-fit test seeks to determine whether the data follows a certain theoretical distribution; the test of independence is used to determine whether two variables have association or are independent; and the test of homogeneity allows to study whether different populations are homogeneous with respect to any variable.
- Be able to select the right test, although he cannot yet develop it.

An approximation of tests with χ^2 statistic

Once the test has been identified, to respond to the problem raised the student must:

- Identify the null hypothesis that is implicit in the problem. Statistical enquiry cycles such as the PPDAC have as their first component the generation of a research question, which must be given in a particular context that is about which one want to inquire. Some investigations (e.g., Pfannkuch & Wild 2004; Pfannkuch et al., 2016; Stohl Lee et al., 2010), have taken up this first component, recognizing that most of these questions have form of conjecture or hypothesis.
- Enunciate and use the properties of the χ^2 distribution (e.g., is positively skewed, has as its only parameter the degrees of freedom, as the degrees of freedom increase it approaches to the normal curve and cannot take negative values). In order for the student to understand the notion of distribution, some research has relied on technology, generating, for example, simulations (e.g., Bakker & Gravemeijer, 2004; Dinov et al., 2018; Reading & Reid, 2006; Rossman, 2008).

The above indicators are for the three tests of the χ^2 (M1, M2, and M3). But, additionally, other indicators to consider by test are:

<u>Goodness-of-fit-test (M1χ^2):</u> The student can assess the extent to which a group of observed data is adjusted to a certain preset theoretical distribution, by contrasting the observed frequencies and theoretical frequencies. To make this assessment, the student:

- Calculates the χ^2 statistic enunciating concepts or definitions and properties or propositions of PM2χ^2 (e.g., observed frequency, theoretical or expected frequency), the Chi-square statistic seen as property $\chi^2 = \sum \frac{e^2}{m} = \sum \frac{(FO-FE)^2}{FE}$ and the error as $e = \mathrm{FO} - FE$.
- Calculates and enunciates the degrees of freedom. According to its definition, indicates that it refers to the number of rows minus the number of independent linear restrictions on the frequencies, and the property $k = n - r$ (PM3χ^2).
- Uses the probability table of the χ^2 distribution to determine probability and can interpret it as a measure of occurrence of a complex system of n errors occurring with a frequency as large or larger than that of the observed system (for which the concept/definition of probability and the probability distribution property χ^2 is used, corresponding to PM2χ^2). According to Stohl Lee et al. (2010), students will be making decisions naturally, either to maintain their current

Table 2. Tabular representation & symbology

a	*b*	*a+b*
c	*d*	*c+d*
a+c	*b+d*	*a+b+c+d*

Table 3. Calculation of expected frequency

(a+b)(a+c)/(a+b+c+d)	*(a+b)(b+d)/(a+b+c+d)*	*a+b*
(c+d)(a+c)/(a+b+c+d)	*(c+d)(b+d)/(a+b+c+d)*	*c+d*
a+c	*b+d*	*a+b+c+d*

hypothesis or to alter it based on the probability obtained.

Inter-level generalization process: Level 1.2, where PM1χ^2 is considered, can only be applied when the purpose is to determine if the dataset follows a normal distribution; however, when working with the elements of level 2.2 (PM2χ^2) the data can be contrasted with any theoretical distribution.

Test of independence (M2χ^2) & homogeneity (M3χ^2): Because the tests of independence and homogeneity are mathematically identical, they share the following indicators:

- Represent the observed frequency with $r \times$ c tables (PM6χ^2 concept). For the specific case of the 2×2 contingency tables, recognize and use the tabular representation and symbology, as shown in **Table 2**.
- Calculate the expected frequency under probabilistic independence (PM6χ^2 concepts and PM11χ^2 properties).
  - For 2×2 contingency tables (**Table 3**).
  - For contingency tables of $r \times$ c, use Eq. (2).

$$e_{ij} = \frac{n_i \times n_j}{n}. \quad (2)$$

  - For $r \times$ c contingency tables, use Eq. (3).

$$\chi^2 = \sum\left\{\frac{(n_{uv} - v_{uv})^2}{v_{uv}}\right\} = \sum \frac{(FO - FE)^2}{FE}. \quad (3)$$

- Enunciate properties of the PM7 to calculate the degrees of freedom $k = (c-1)(r-1)$.

Intra-level generalization process: At level 2.2, when calculating the expected frequencies there is a generalization process, as well as when calculating the χ^2 statistic, from the 2×2 contingency tables to those of $r \times c$.

Test of independence: The student can determine whether there is association between two variables (with n attributes each variable), for which he assesses to what extent the observed frequencies differ from probabilistic independence. To determine such an association, primary mathematical objects of PM6χ^2 (e.g., variable concepts, contingency table, association) intervene. Once the student has calculated the value of the statistic and the degrees of freedom:

- Identifies and enunciates the use that χ^2 statistic has for the test of independence.
- Uses the probability table of the χ^2 distribution to determine probability and can interpret it as a measure of how far the observed system is compatible with the probabilistic independence bases.

Inter-level generalization process: Level 1.3, where the PM5χ^2 is used, can only be applied when the objective is to know if there is association between two variables, with one attribute each variable. However, when working with the elements of level 2.2, PM6χ^2, the association between two variables with n attributes each variable can be analyzed.

Test of homogeneity: The student can determine whether two known samples may be from the same population, without having a priori knowledge of the population. Some of the primary mathematical objects of the PM9χ^2 involved are sample concepts, independent samples, and population. Furthermore:

- Identifies and enunciates the use that χ^2 statistic has for homogeneity.
- Uses the probability table of the χ^2 distribution to determine the probability and can interpret it as a measure of the samples actually being random samples of the same population.

Level 3–Pre-Formal

Indicators at this level can be considered pre-formal, but with a higher degree of formality than at level 2. The aspects that mark a certain pre-formal degree are, for example, the way of working and arguing about the significance and the language used in the hypotheses.

Restrictions of χ^2 tests

The student can recognize the restrictions of the hypothesis tests with the χ^2 statistic and apply the continuity correction factor when necessary.

- Uses the χ^2 distribution as an approximation. This is because the χ^2 distribution is continuous, while the distribution we are trying to approximate is discrete.
- Identifies and enunciates the limitations that this causes in the test.
- Recognizes when to apply the continuity correction factor (PM4χ^2 concept)
- Enunciates the null and alternative hypothesis in natural language.

In addition, some indicators per test are proposed:

Goodness-of-fit-test:

- Defines continuity correction factor and enunciates the following property to calculate it (concepts and properties of PM4χ^2), as shown in Eq. (4).

$$\chi_c^2 = \sum \frac{(|e|-0.5)^2}{m} = \sum \frac{(|FO-FE|-0.5)^2}{FE}. \quad (4)$$

Test of independence and homogeneity:

- Defines the continuity correction factor and enunciates the following property to calculate it (concept of PM4χ^2 and property of PM8χ^2):
  - For 2×2 contingency tables, use Eq. (5):

$$\sum \chi_c^2 = \sum \frac{\left(\left(a-\frac{1}{2}\right)\left(d-\frac{1}{2}\right)-\left(b+\frac{1}{2}\right)\left(c+\frac{1}{2}\right)\right)^2 N}{(a+c)(b+d)(a+b)(c+d)}. \quad (5)$$

  - For $r \times$ c contingency table, use Eq. (6):

$$\chi_c^2 = \sum \frac{(|FO-FE|-0.5)^2}{FE}. \quad (6)$$

Intra-level generalization process: At level 3.1, when calculating the value of the χ^2 statistic, there is a generalization process, the way to calculate it when working with 2×2 contingency tables to those of $r \times c$.

Connections & arguments

- Defines significance as indicative of the level at which the possibility of the effect should receive serious consideration (PM2χ^2 definition of significance).
- Can find the value of the theoretical statistic, in the probability table of the χ^2 distribution, with respect to certain P and n; and compares it against the value of the calculated χ^2 statistic.
- Is capable of rejecting or not rejecting the null hypothesis under a contrast with a pre-established limit as a significant deviation, according to the decision rules regarding the probability and the value of the statistic (properties/propositions of the PM3χ^2 and PM4χ^2, respectively).
- Manages to argue, based on significance, why he rejects or does not reject the null hypothesis.
- Can connect test results to the context of the problem.

In the indicators of levels two and three we can find the three key principles (generalization, use of data as evidence and the use of probabilistic language) that Makar and Rubin (2009) indicate as essential for informal statistical inference. These three key principles are found in different depths at both levels, since as they point out in this framework for informal inference, these principles can be applied at different moments of the school curriculum with different depths.

Level 4–Formal

In this level, indicators that correspond to a FIR are presented. As it was worked in previous levels and as it was observed in history, the hypothesis test is first addressed with the *p-value* and then with the critical region. The last indicators correspond to working with error type I and II, and the power of the test.

Criteria for decision-making

It is expected that the student could make decisions based on the statistical techniques of the hypothesis tests methodology.

p-value:

- Can state the null and alternative hypothesis with symbolic and natural language.
- Identifies and uses the common values of the significance level. For example, 0.05 is very common and was popularized in Fisher's test; and 0.10, and 0.01 are used as significance level, which selection depends on the magnitude of the error that it is wanted to assume. It is also important to define (and use in the practice) the value of the significance level as the probability of concluding that a deviation or difference exists when in fact it does not exist; in other words, it is the probability of failing in the estimation.
- Identifies the relation between significance level (α) and confidence level ($1-\alpha$).
- Enunciates and is able to apply the decision rule. If the $value - p < \alpha$, is rejected H_0, (property/proposition of PM3χ^2).
- Interprets the p-value as a continuous probability measure that the value of the calculated statistic is possible given the null hypothesis, with which it is usually interpreted that when there is a high p-value the null hypothesis is not rejected, although when the p-value is low it is rejected. How low? Depends on the level of significance (Cohen, 1994). However, we have to take caution because, on the one hand, the p-value and the level of significance do not quantify the effect size or the importance of a result and, on the other hand, the p-value does not measure the probability that the null hypothesis is true or that the alternative hypothesis is correct.

Critical value:

- Can identify the theoretical value of the χ^2 statistic, according to the significance level and the degrees of freedom.
- Can graphically represent the regions of acceptance and rejection and can argue about the relationships with the confidence level and the null hypothesis.
- Enunciates and is able to apply the decision rule. If $\chi^2 \geq \chi^2_{\alpha,gl}$, H_0 is rejected (property/proposition of PM4χ^2).
- Can provide a response to the problem by using the results of the test and develops arguments with statistical basis.

Interpretation together with arguments, evidences the understanding and connections the student makes of statistical notions (Lane-Getaz, 2013; Makar et al., 2011).

Error type I & II, & power of test

The student can evaluate the validity of the procedures and the inferences made based on such procedures, for which:

- Argues when the error type I is made and the probability of making it: $P\begin{bmatrix}Reject\ H_0|H_0\\ is\ true\end{bmatrix} = \alpha$.
- Argues when the error type II is made and the probability of making it: $P\begin{bmatrix}Not\ reject\ H_0|H_0\\ is\ false\end{bmatrix} = \beta$.
- Can calculate the probability of making the correct decision when H_0 is true and when H_0 is false.
- Identifies the relation between error type I and the error type II. For example, can use a graphical representation and recognize that the larger α is, the smaller β is. It is also important not to confuse the conditional probabilities involved in type I and type II errors with single event probabilities (Birnbaum, 1982; Falk, 1986; Shaughnessy & Dick, 1991; Sotos et al., 2007).
- Enunciates the power of the test and can calculate it. Recognizes that power is related to sample size and significance level: $P\begin{bmatrix}Decide\ H_1|H_1\\ is\ true\end{bmatrix} = 1-\beta$.
- Argues about the validity of the inference made.

According to Wild et al. (2018), in addition to the technical aspects, the reasoning behind the hypothesis tests is very important, since it can even be used to make decisions without quantitative data. Likewise, understanding the importance of expected values can help us make better decisions even in our daily lives.

Figure 3 shows a summary of the indicators of the inferential reasoning levels for the Chi-square statistic,

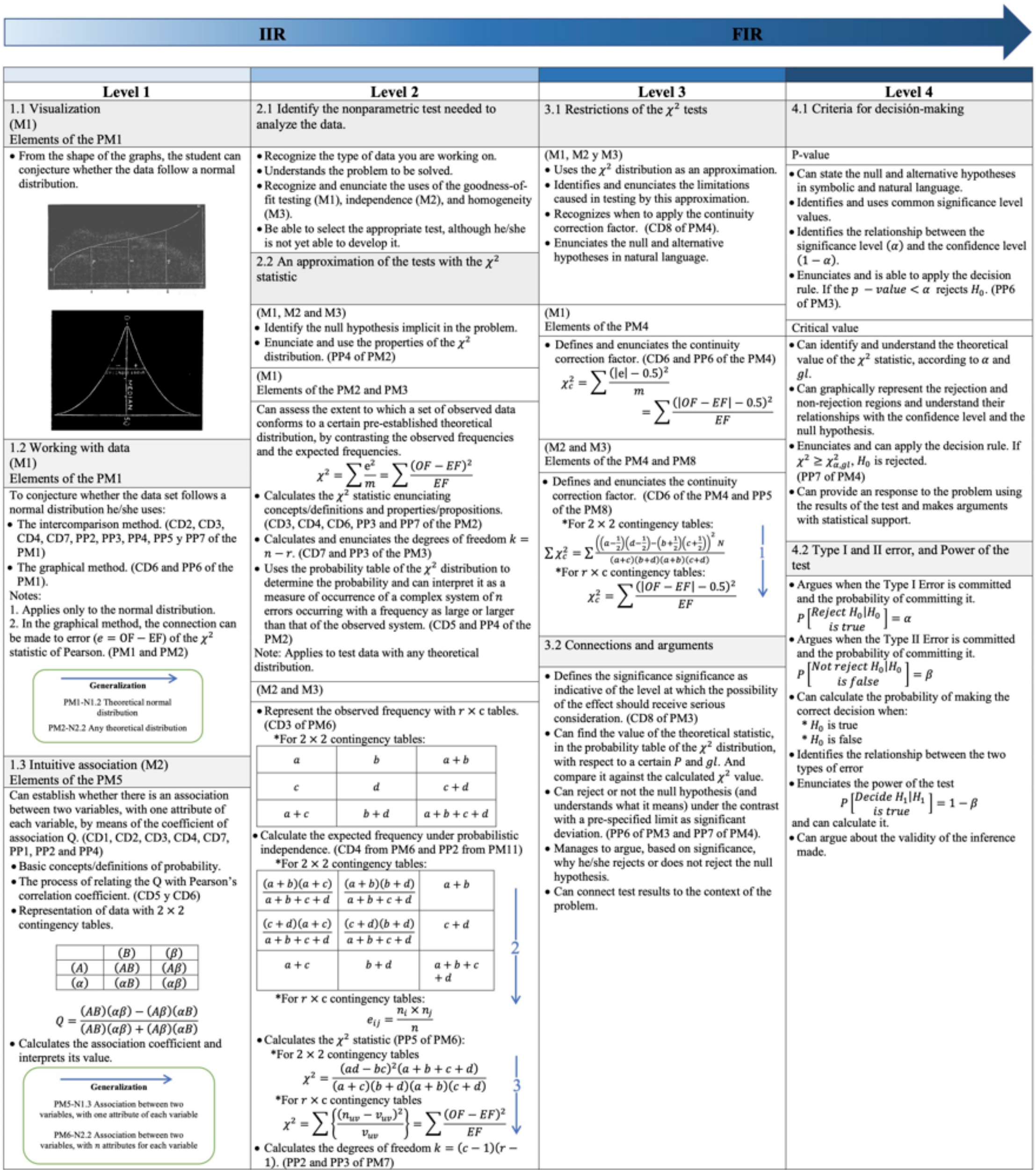


Figure 3. Progressive levels of inferential reasoning (from informal to formal) (Source: Authors' own elaboration)

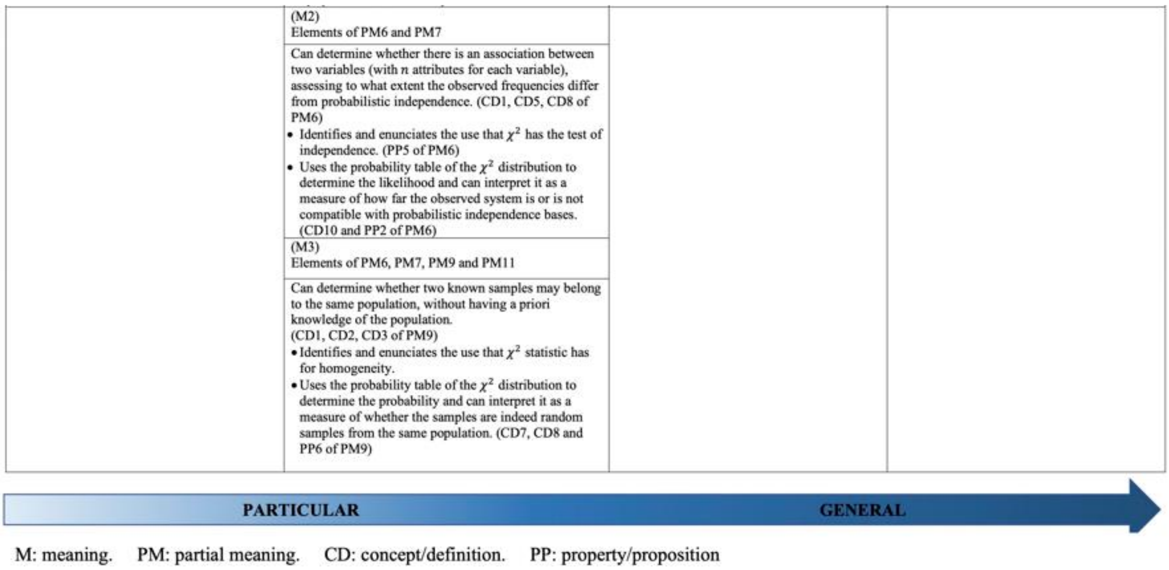


Figure 3 (continued). Progressive levels of inferential reasoning (from informal to formal) (Source: Authors' own elaboration)

which are gradual and progressive in terms of generalization and formalization.

INFERENTIAL REASONING ON THE CHI-SQUARE STATISTIC

To explain the inferential reasoning that could be promoted with the proposed levels of inferential reasoning for the Chi-squared statistic, we present examples of practices that "activate" each level. It should be noted that other finer analyses can be developed, and we do not intend to be exhaustive in the possible answers and their analyses, we only give a look at the use of the proposal of levels. For this we have adapted a Pearson (1904) problem, which allows us to transit on all levels.

Situation/Problem

Table 4 shows the data from a sample that was collected during a smallpox epidemic that occurred in 1890 in a small town. The data in Table 4 concern the presence or absence of the scar of the smallpox vaccine and whether the people who received the vaccine recovered or died. Is there a link between the presence of vaccine scarring and smallpox recoveries?

Example of Response Associated with Level 1

Initially, if we look at the data from the 2×2 contingency table, we can observe a large concentration at the intersection of those who have a scar and those who recovered, which may lead us to think that there may be an association between the recovery and the presence of the scar. If we analyze the data in the same sense as the correlation coefficient, but now with discrete variables using the coefficient of association, to test if there is an association between the variables. We calculate Q using Eq. (1), as follows:

Table 4. Presence of scar from vaccine & recovered from smallpox

Scar	Type of response		
	Recoveries	Deaths	Totals
Present	35	1	36
Absent	10	4	14
Totals	45	5	50

$$Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)} = \frac{(35)(4)-(1)(10)}{(35)(4)+(1)(10)} = \frac{140-10}{140+10} = 0.8666.$$

According to the value of the coefficient of association that is 0.8666, we can say that there is a relationship between the variables, that is, there is a relationship between the presence of a scar from the vaccine and those recovered from smallpox.

Example of Response Associated with Level 2

We are working with a sample of 50 individuals who received a smallpox vaccine and are classified into two (discrete) variables. One of the variables refers to a side effect of the vaccine, which is a scar, having as attributes the "presence or absence" of the scar left by the vaccine, while the attributes of the second variable are "recovered or dead". Due to the type of data presented and since it is desired to know if there is any relationship between the presence of a scar from the vaccine and those recovered from smallpox, which seems to be the hypothesis that exists, a test of independence can be applied with the Chi-square statistic to test whether or not there is independence between variables.

If we apply the test of independence, we can calculate the expected frequencies under independence to later calculate the Chi-square statistic, but since it is a 2×2

contingency table, we can apply Eq. (3): $\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+c)(b+d)(a+b)(c+d)} = \frac{((35\times4)-(1\times10))^2(50)}{(35+10)(1+4)(35+1)(10+4)} = 7.45149$.

Now, we proceed to calculate the degrees of freedom under the formula $k = (c-1)(r-1) = (2-1)(2-1) = 1$.

From these values and through the Chi-square distribution, we can obtain the probability that the data in the table are compatible with the probabilistic independence bases. There is a probability of occurrence of values such as those observed of 0.006338, that is, we could only see values like these 6.33 times in 1,000 cases if the variables were independent.

From the procedures carried out with the test of independence, we could say that there is no independence between the presence of a scar from the vaccine and those recovered from smallpox.

We can also consider the option of the student using statistical software to analyze the data. It is important that he can identify (on the output screen) and interpret the above. He could even establish the differences between the expected and observed frequencies as aspects of interest. It is important to note that in this case the student would ignore the warning generated by the software for observed frequencies below five.

Example of Response Associated with Level 3

To exemplify a response at this level, consider what was outlined in the response example associated with level two and the following:

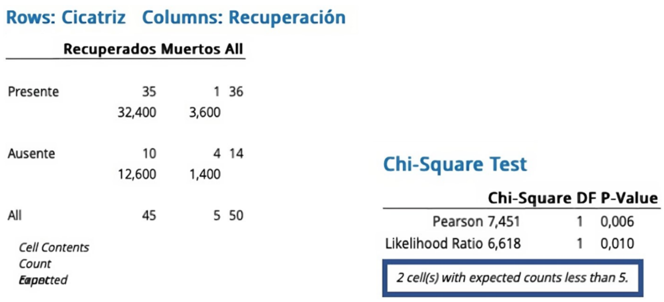
The null and alternative hypotheses are,

Ho: $Variables\ are\ independent$ (there is no association between scar and smallpox recovery).

Ha: $There\ is\ no\ independence\ between\ the\ variables$ (there is an association between the scar and the recovery of smallpox). Two boxes in the 2×2 contingency table have observed frequencies below five, whence we must take precautions. This is because when we use the test of independence with small numbers, we can obtain a discrepancy because the test is made with a continuous distribution, while the distribution that is intended to approximate is direct. To reduce this discrepancy, we can use the Yates continuity correction factor. Then we calculate the statistic under Eq. (4): $\chi_c^2 = \frac{\left(\left(a-\frac{1}{2}\right)\left(d-\frac{1}{2}\right)-\left(b+\frac{1}{2}\right)\left(c+\frac{1}{2}\right)\right)^2 N}{(a+c)(b+d)(a+b)(c+d)} = \frac{\left(\left(35-\frac{1}{2}\right)\left(4-\frac{1}{2}\right)-\left(1+\frac{1}{2}\right)\left(10+\frac{1}{2}\right)\right)^2 50}{(35+10)(1+4)(35+1)(10+4)} = \frac{(120.75-15.75)^2(50)}{(45)(5)(36)(14)} = 4.86111$.

With 4.86111 as the value of the statistic and with one degree of freedom we have a probability of occurrence of having values, under independence, such as those observed of 0.02746.

If we consider a probability of 0.05 as a significant deviation limit, since the probability obtained is less, we will have to say that the variables are not independent.


Rows: Cicatriz Columns: Recuperación

	Recuperados	Muertos	All
Presente	35	1	36
	32,400	3,600	
Ausente	10	4	14
	12,600	1,400	
All	45	5	50

Cell Contents
Count
Expected

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	7,451	1	0,006
Likelihood Ratio	6,618	1	0,010

2 cell(s) with expected counts less than 5.

Figure 4. Output screen of test for association in Minitab (Source: Authors' own elaboration)

Under this same limit the theoretical statistic is 3.84145. And if we compare them, the calculated statistic is larger than the theoretical statistic. This means that it exceeds our limit, therefore the deviations from expectations are clearly significant.

Another possibility that is within this level, although to a lesser degree, would be that the student uses software to take the test and that, when seeing the warning on the output screen - which in this case would indicate that "two cells have a count less than 5 "– as highlighted in **Figure 4**, the student develops a reflection of how frequencies below five could affect the value of the statistic and therefore the probability, and why. However, the student would not take the test with the continuity correction factor.

Example of Response Associated with Level 4

In addition to what was expressed in the answer of the previous level, an answer at this level could consider the following aspects:

The problem it poses does not indicate the significance level, whence I will use the most common one, $\alpha = 0.05$.

Considering that we have a statistic value with correction $\chi_c^2 = 4.86111$ and with a p-value of 0.02746, we can see that our p-value is less than alpha, thence Ho is rejected. Another criterion for making the decision is to contrast the value of the statistic that we calculate with the theoretical one and as $\chi^2_{.05,1} = 3.84145$, we have then that $4.86111 > 3.84145$, therefore, the difference is significant, which leads us to reject Ho.

In **Figure 5** we can see that the graph on the left shows the critical region from the value of the theoretical statistic (according to the significance level of and degrees of freedom), while in the graph on the right we can see the statistic calculated and its associated probability, and how it is within the rejection zone.

Based on the above, with a confidence level of 95% we can reject Ho and accept Ha as probably true, that is, that the variables are not independent and therefore there is an association between scar and smallpox recovery.

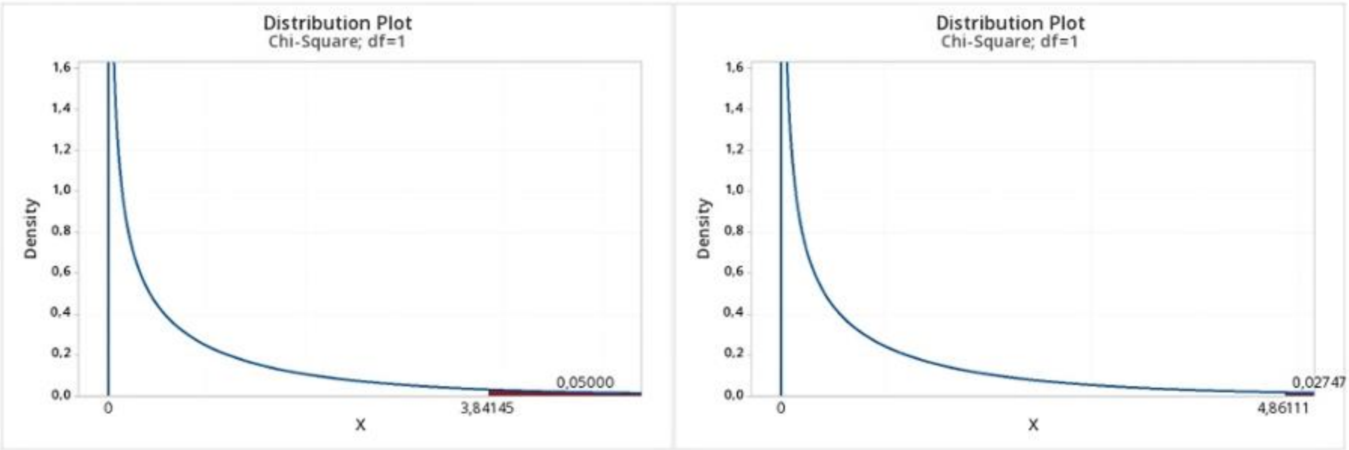
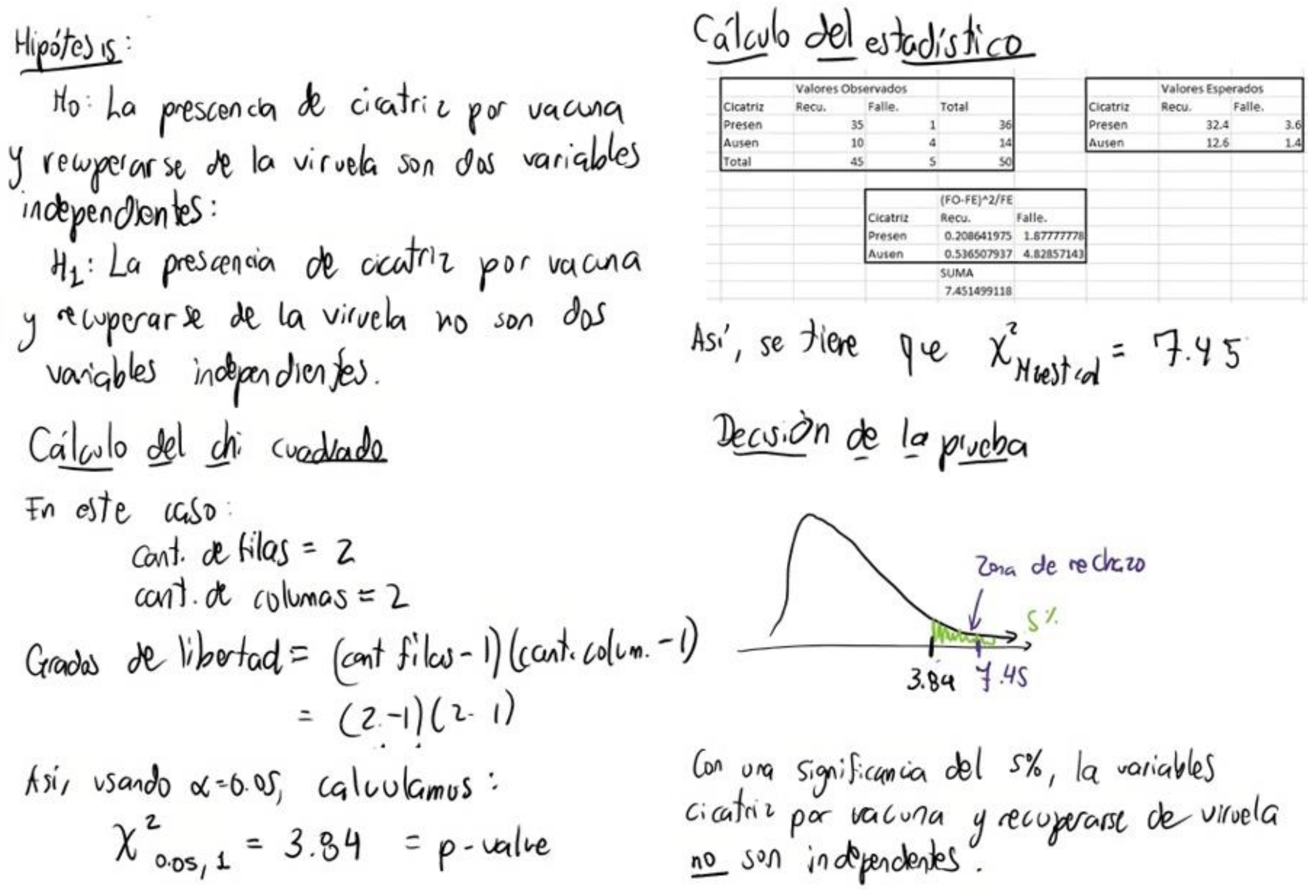


Figure 5. Chi-square distribution graphs for values of statistic of 3.84145 & 4.86111 (Source: Authors' own elaboration)


Hipótesis:

Ho: La prescencia de cicatriz por vacuna y recuperarse de la viruela son dos variables independientes:

H_1: La prescencia de cicatriz por vacuna y recuperarse de la viruela no son dos variables independientes.

Cálculo del chi cuadrado

En este caso:

cant. de filas = 2

cant. de columnas = 2

Grados de libertad = (cant filas − 1)(cant. colum. − 1)

= (2 − 1)(2 − 1)

Así, usando $\alpha = 0.05$, calculamos:

$\chi^2_{0.05,1} = 3.84$ = p-valve

Cálculo del estadístico

Valores Observados			
Cicatriz	Recu.	Falle.	Total
Presen	35	1	36
Ausen	10	4	14
Total	45	5	50

Valores Esperados		
Cicatriz	Recu.	Falle.
Presen	32.4	3.6
Ausen	12.6	1.4

(FO-FE)^2/FE		
Cicatriz	Recu.	Falle.
Presen	0.208641975	1.87777778
Ausen	0.536507937	4.82857143
	SUMA	
	7.451499118	

Así, se tiene que $\chi^2_{Muestral} = 7.45$

Decisión de la prueba

Con una significancia del 5%, la variables cicatriz por vacuna y recuperarse de viruela no son independientes.

Figure 6. Practice of prospective teacher (Source: Authors' own elaboration)

The probability that we have of making the error type I, since we have rejected Ho, is 0.05, that is, of rejecting Ho when it is really true. Regarding the power of the test, a minimum of 80% is usually requested, and in this case, using software to perform the calculation, we obtained a probability of 0.86275, which indicates that there is a constant prevalence in the population. We can also say that the probability of making the error type II in this test is 0.13725.

Consequently, in **Figure 6** we present an example of a response made by a prospective teacher on this same problem.

However, although the prospective teacher mobilizes mathematical objects that correspond to level 4, he seems to misunderstand the p-value, since he is really working with the critical value or statistical value. The proposal of levels of inferential reasoning on the Chi-square statistic promotes the progressive understanding of notions such as the p-value. The prospective teacher could start with an informal approach to the p-value, using software for simulations or for calculating the probability of the Chi-square statistic (level 2), and thus support his or her inference (Rossman, 2008; Rossman & Chance, 2014). For the teacher to develop a more robust level 4 practice, he could work with type I and type II errors and understand the relationships between these errors and validate his inference by making use of the power of the test.

FINAL REFLECTIONS

The aim of this paper was to present a proposal of progressive levels of inferential reasoning, from informal to formal, on the Chi-square statistic, based on the mathematical richness retrieved from the historical-epistemological study on this statistic and from the statistics education literature on inferential reasoning. This proposal is not intended to be definite; on the contrary, we consider it to be an initial approach to levels of inferential reasoning. Certainly, it could be extended to other statistics, and from hypothesis testing to confidence intervals, to name a few. This will occur as additional studies of this type, and historical-epistemological studies on the key notions of inference, are developed.

The proposal consists of four progressive levels, the indicators of the first level are closely linked to an *IIR*, while those of the fourth level to a *FIR*. Therefore, both the second and the third level, which we have called pre-formal, contain features of both informal and formal inference, though at varying degrees of graduality. Thus, the four levels provide gradable 'indicators' for generalization and formalization processes. We consider that this approach allows us to account for a seamless transition from an IIR to a FIR, in other words, it could provide a mechanism to encourage students to develop an IIR first, and then build a FIR based on that reasoning.

An important element to note is that although we consider visualization as a key aspect at level one, it is not exclusive to it. It is initially desirable that students could make conjectures only from the information presented in the graph through the elements presented in it (see level 1), which would be consistent with various proposals on IIR (e.g., Zieffler et al., 2008). On other levels, visualization can be used for other purposes, for example, to understand the distribution (level 2), and to graphically represent the acceptance and rejection regions for the null hypothesis (level 4). These levels are related to the mathematical practices that are developed to solve a problem, that is, the levels are not predictors of problem types but of mathematical practice. Then, the same problem can activate practices associated with any of the four levels that we propose.

Although the levels are progressive, it is not necessary for the student to transit through the four of them. The teacher can decide whether the class should be based on the criteria of level 1 or another level. However, we suggest that the transition should be made through the four levels of inferential reasoning proposed, as they allow the student to approach the mathematical object intuitively at first, then pre-formally and lastly formally. Therefore, we can say that the criteria for each level bring the student closer to the criteria of the next level, which is consistent with statistics education research studies (e.g., Makar & Rubin, 2018; Pfannkuch et al., 2015; Zieffler et al., 2008).

It should be noted that the proposal's perspective on *reasoning* is based on a pragmatist view of mathematical knowledge formation (as well as school mathematical knowledge), which incorporates the semiotic, anthropological, and pragmatic postulates of OSA (Godino et al., 2007, 2019). That is to say, in accordance with the suggestions of Aké (2013) for algebra, and Molina (2019) for geometry, in order to speak about school statistics, one must resort to an integrated and transdisciplinary vision that involves Statistical thinking, Statistical reasoning, and Statistical literacy, since each of these approaches to school statistics is developed from psychological, epistemic, and semiotic perspectives, respectively. Thus, by considering reasoning in terms of practices, and (primary mathematical) objects and mathematical processes used in them, one moves in a certain sense through the definitions given by Ben-Zvi and Garfield (2004) for statistical literacy, statistical reasoning and statistical thinking.

The *indicators* of the IR levels here proposed, can serve as an initial guideline for lesson planning, designing activities that promote inferential reasoning progressively, and studying the levels of IR being promoted in the mathematical practice of students, teachers, or the curriculum, as observed in previous studies (Lugo-Armenta & Pino-Fan, 2021a, 2022). Teaching based on the indicators of our proposed IR levels could have an impact on students' understanding of the Chi-square statistic and related notions in their last pre-university and first university years. For example, we can refer to a progression to work at various times on the p-value (levels 2, 3, and 4), significance (level 3 and level 4) and the posing of the null and alternative hypotheses (levels 2, 3, and 4); notions that might lead to errors and generate difficulties for students and teachers, as observed in previous studies (Biehler et al., 2015; López-Martín et al., 2019; Sotos et al., 2007; Vera & Díaz, 2013).

It is noteworthy that currently, there is an ongoing discussion about the p-value and the level of significance, initially promoted by the American Statistical Association (ASA) in ASA statement on p-values and statistical significances, where they recommended abandoning the declaration of 'statistical significance'. This recommendation stems from the usage observed both in the teaching of statistics and in published research, where the p-value has often been perceived more as a static rule rather than considering it as a continuous probability value. We hold that the indicators of the proposed levels of IR, considering the progression of the p-value and the level of significance at various levels could assist in enabling students to interpret the p-value appropriately and to recapture Fisher's (1925) original concept that statistical significance should be viewed as a tool indicating when a result, 'deemed significant', warrants further scrutiny.

The nature of the proposed IR levels' indicators transcends the purely algorithmic aspects and focuses on the progression of inferential reasoning using various mediational resources, hence these proposed levels can be used to promote or characterize inferential reasoning, whether with or without the use of statistical software, this arises from algorithmic developments or internal processes of the software alone are insufficient for making inferences. In both cases it becomes crucial, besides the conclusion to the problem within its own context, the evidence and reasoning upon the inference was realized, aligns with Rossman's (2008) proposal regarding what constitutes inference. The proposal of levels of inferential reasoning on the Chi-square statistic has a progressive nature, allowing for exploration of the various notions at various moments (in the levels) with varying degrees of complexity, depth, and formality.

Author contributions: All authors have sufficiently contributed to the study and agreed with the results and conclusions.

Funding: This study was developed within the framework of research project Fondecyt 1200005 funded by Agencia Nacional de Investigación y Desarrollo (ANID) of Chile and the research project I+D FIED21-002 funded by Secretaría Nacional de Ciencia, Tecnología e Innovación (SENACYT) / Universidad de Panamá (Panamá).

Ethical statement: The authors stated that the study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Universidad de Los Lagos (Folio CEC-ULagos H012/2020 and date of approval 7 July 2020).

Declaration of interest: No conflict of interest is declared by authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Aké, L. P. (2013). *Evaluación y desarrollo del razonamiento algebraico elemental en maestros en formación* [*Evaluation and development of elementary algebraic reasoning in training teachers*] [Doctoral dissertation, Universidad de Granada].

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning, 13*(1-2), 5-26. https://doi.org/10.1080/10986065.2011.538293

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147-168). Springer. https://doi.org/10.1007/1-4020-2278-6_7

Bakker, A., Ben-Zvi, D., & Makar, K. (2017). An inferentialist perspective on the coordination of actions and reasons involved in making a statistical inference. *Mathematics Education Research Journal, 29*(4), 455-470. https://doi.org/10.1007/s13394-016-0187-x

Batanero, C. (2013). Del análisis de datos a la inferencia: Reflexiones sobre la formación del razonamiento estadístico [From data analysis to inference: Reflections on the formation of statistical reasoning]. *Cuadernos de Investigación y Formación en Educación Matemática* [*Research and Training Notebooks in Mathematics Education*], *11*, 277-291.

Batanero, C., Vera, O. D., & Díaz, C. (2012). Dificultades de estudiantes de psicología en la comprensión del contraste de hipótesis [Difficulties of psychology students in understanding the contrast of hypotheses]. *Números. Revista de Didáctica de las Matemáticas* [*Numbers. Mathematics Didactics Magazine*], *80*, 91-101.

Ben-Zvi, D., & Aridor-Berger, K. (2016). Children's wonder how to wander between data and context. In D. Ben-Zvi, & K. Makar (Eds.), *The teaching and learning of statistics: International perspectives* (pp. 25–36). Springer. https://doi.org/10.1007/978-3-319-23470-0_3

Ben-Zvi, D., & Garfield, J. B. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). Springer. https://doi.org/10.1007/1-4020-2278-6

Biehler, R., Frischemeier, D., & Podworny, S. (2015). Preservice teachers reasoning about uncertainty in the context of randomization tests. In A. Zieffler, & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning* (pp. 129-162). Catalyst Press.

Cañadas, G., Batanero, C., Díaz, C., & Gea, M. M. (2012). Comprensión del test Chi-cuadrado por estudiantes de psicología [Understanding of the Chi-square test by students of psychology]. In A. Estepa, Á. Contreras, J. Deulofeu, M. C. Penalva, F. J. García, & L. Ordóñez (Eds.), *Investigación en educación matemática XVI* [*Research in mathematics education XVI*] (pp. 153 - 163). SEIEM.

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist, 49*(12), 997-1003. https://doi.org/10.1037/0003-066X.49.12.997

Dierdorp, A., Bakker, A., Eijkelhof, H., & van Maanen, J. (2011). Authentic practices as contexts for learning to draw inferences beyond correlated data. *Mathematical Thinking and Learning, 13*(1-2), 132-151. https://doi.org/10.1080/10986065.2011.538294

Dinov, I. D., Palanimalai, S., Khare, A., & Christou, N. (2018). Randomization-based statistical inference: A resampling and simulation infrastructure.

Teaching Statistics, 40(2), 64-73. https://doi.org/10.1111/test.12156

Doerr, H. M., delMas, R. & Makar, K. (2017). A modeling approach to the development of students' informal inferential reasoning. *Statistics Education Research Journal, 16*(2), 86-115. https://doi.org/10.52041/serj.v16i2.186

Dolor, J., & Noll, J. (2015). Using guided reinvention to develop teachers' understanding of hypothesis testing concepts. *Statistics Education Research Journal, 14*(1), 60-89. https://doi.org/10.52041/serj.v14i1.269

English, L. D., & Watson, J. (2018). Modelling with authentic data in sixth grade. *ZDM-International Journal on Mathematics Education, 50*(1-2), 103-115. https://doi.org/10.1007/s11858-017-0896-y

Estrella, S., Méndez-Reina, M., Salinas, R., & Rojas, T. (2023). The mystery of the black box: An experience of informal inferential reasoning. In G. F. Burrill, L. de Oliveria Souza, & E. Reston (Eds.), *Research on reasoning with data and statistical thinking: International perspectives* (pp. 191-210). Springer. https://doi.org/10.1007/978-3-031-29459-4_16

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.

Font, V., & Rubio, N. V. (2017). Procesos matemáticos en el enfoque onto-semiótico [Mathematical processes in the onto-semiotic approach]. In J. M. Contreras, P. Arteaga, G. R. Cañadas, M. M. Gea, B. Giacomone, & M. M. López-Martín (Eds.), *Actas del Segundo Congreso International Virtual sobre el Enfoque Ontosemiótico del Conocimiento y la Instrucción Matemáticos* [*Proceedings of the 2nd International Virtual Congress on the Ontosemiotic Approach to Mathematical Knowledge and Instruction*]. Universidad de Granada.

Font, V., Godino, J. D., & Gallardo, J. (2013). The emergence of objects from mathematical practices. *Educational Studies in Mathematics, 82*(1), 97-124. https://doi.org/10.1007/s10649-012-9411-0

Galton, F. (1875). IV. Statistics by intercomparison, with remarks on the law of frequency of error. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 4, 49*(322), 33-46. https://doi.org/10.1080/14786447508641172

Galton, F. (1885). The application of a graphic method to fallible measures. *Journal of the Statistical Society of London*, 262-271.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer. https://doi.org/10.1007/978-1-4020-8383-9

Gil, E., & Ben-Zvi, D. (2011). Explanations and context in the emergence of students' informal inferential reasoning. *Mathematical Thinking and Learning, 13*(1-2), 87-108. https://doi.org/10.1080/10986065.2011.538295

Godino, J. D. (2022). Emergencia, estado actual y perspectivas del enfoque ontosemiótico en educación matemática [Emergence, current state and perspectives of the onto-semiotic approach in mathematics education]. *Revista Venezolana de Investigación en Educación Matemática* [*Venezuelan Journal of Research in Mathematics Education*], *2*(2), e202201. https://doi.org/10.54541/reviem.v2i2.25

Godino, J. D. Batanero, C., & Font, V. (2007). The onto-semiotic approach to research in mathematics education. *ZDM Mathematics Education, 39*(1-2), 127-135. https://doi.org/10.1007/s11858-006-0004-1

Godino, J. D., & Batanero, C. (1994). Significado institucional y personal de los objetos matemáticos [Institutional and personal meaning of mathematical objects]. *Recherches en Didactique des Mathématiques* [*Research in Mathematics Didactics*], *14*(3), 325-355.

Godino, J. D., Batanero, C., & Font, V. (2019). The onto-semiotic approach: Implications for the prescriptive character of didactics. *For the Learning of Mathematics, 39*(1), 37- 42.

Godino, J. D., Font, V., Wilhelmi, M. R. & Lurduy, O. (2011). Why is the learning of elementary arithmetic concepts difficult? Semiotic tools for understanding the nature of mathematical objects. *Educational Studies in Mathematics, 77*(2), 247-265. https://doi.org/10.1007/s10649-010-9278-x

Godino, J. D., Neto, T., Wilhelmi, M., Aké, L., Etchegaray, S., & Lasa, A. (2015). Algebraic reasoning levels in primary and secondary education. In K. Krainer, & N. Vondrová, (Eds.) *Proceedings of 9th Congress of the European Society for Research in Mathematics Education* (pp. 426-432). ERME.

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). Springer. https://doi.org/10.1007/978-94-007-1131-0_24

Jacob, B. L., & Doerr, H. M. (2014). Statistical reasoning with the sampling distribution. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Proceedings of the 9th International Conference on Teaching Statistics*. IASE.

López-Martín, M. D. M., Batanero, C., & Gea, M. M. (2019). ¿Conocen los futuros profesores los errores de sus estudiantes en la inferencia estadística? [Do prospective teachers know their students' errors in statistical inference?] *Bolema: Boletim de Educação Matemática* [*Bulletin: Mathematics Education*

Bulletin], *33*(64), 672-693. https://doi.org/10.1590/1980-4415v33n64a11

Lugo-Armenta, J. G., Pino-Fan, L. R., & Ruiz, B. (2021). Meanings of the Chi-square statistic: A historical-epistemological overview. *Revemop - Revista de Educação Matemática de Ouro Preto, 3*, 1-33. https://doi.org/10.33532/revemop.e202108

Lugo-Armenta, J. G., & Pino-Fan, L. R. (2021a). Inferential reasoning of high school mathematics teachers about chi-squared statistic. *Mathematics, 9*(19), 2416. https://doi.org/10.3390/math9192416

Lugo-Armenta, J. G., & Pino-Fan, L. R. (2021b). Niveles de razonamiento inferencial para el estadístico t-student [Levels of inferential reasoning for the t-student statistician]. *Bolema: Boletim de Educação Matemática, 35*, 1776-1802. https://doi.org/10.1590/1980-4415v35n71a25

Lugo-Armenta, J. G., & Pino-Fan, L. R. (2022). Razonamiento inferencial de profesores de matemáticas de enseñanza media sobre el estadístico t-student [Inferential reasoning of high school mathematics teachers on the t-student statistic]. *Uniciencia, 36*(1), 1-29. https://doi.org/10.15359/ru.36-1.25

Makar, K. (2016). Developing young children's emergent inferential practices in statistics. *Mathematical Thinking and Learning, 18*(1), 1-24. https://doi.org/10.1080/10986065.2016.1107820

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. https://doi.org/10.1080/10986065.2011.538291

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82-105. https://doi.org/10.52041/serj.v8i1.457

Makar, K., & Rubin, A. (2018). Learning about statistical inference. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 261-294). Springer. https://doi.org/10.1007/978-3-319-66195-7_8

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1-2), 152-173. https://doi.org/10.1080/10986065.2011.538301

Matis, T., Riley, L., & Matis, J. (2004). Integrating technologically-based laboratory modules into the stochastic processes curriculum. In G. Burrill, & M. Camden (Eds.), *Curriculum development in statistics education: International association for statistics education* (pp. 93-103). International Statistical Institute. https://doi.org/10.52041/SRAP.04202

Medrano, I., & Pino-Fan, L. R. (2016). Estadios de comprensión de la noción matemática de límite finito desde el punto de vista histórico [Stages of understanding the mathematical notion of finite limit from a historical point of view]. *REDIMAT, Journal of Research in Mathematics Education, 5*(3), 287 - 323. https://doi.org/10.17583/redimat.2016.1854

Mineduc. (2019). *Bases curriculares 3º y 4º medio* [*Curriculum bases 3rd and 4th year*]. Unidad de Currículum y Evaluación. Ministerio de Educación de Chile [Curriculum and Evaluation Unit. Ministry of Education of Chile].

Molina, O. J. (2019). *Sistema de normas que influyen en procesos de argumentación: Un curso de geometría del espacio como escenario de investigación* [*System of norms that influence on argumentation processes: A space geometry course as research scenario*] [Doctoral dissertation, Universidad de Los Lagos].

Ortiz, C. V., & Alsina, Á. (2019). Intuitive ideas about chance and probability in children from 4 to 6 years old. *Acta Scientiae* [*Journal of Science*], *21*(3), 131-154. https://doi.org/10.17648/acta.scientiae.v21iss3id5215

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50*(302), 157-175. https://doi.org/10.1080/14786440009463897

Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. *Draper's Company Research Memoirs Biometric Series I*.

Peirce, C. S. (1958). *Collected papers of Charles Sanders Peirce. 1931-1935*. Harvard UP.

Pfannkuch, M. (2007). Year 11 students' informal inferential reasoning: A case study about the interpretation of box plots. *International Electronic Journal of Mathematical Education, 2*(3), 149-167. https://doi.org/10.29333/iejme/181

Pfannkuch, M., Arnold, P., & Wild, C. J. (2015). What I see is not quite the way it really is: Students' emergent reasoning about sampling variability. *Educational Studies in Mathematics, 88*(3), 343-360. https://doi.org/10.1007/s10649-014-9539-1

Pfannkuch, M., Budgett, S., Fewster, R., Fitch, M., Pattenwise, S., Wild, C., & Ziedins, I. (2016). Probability modeling and thinking: What can we learn from practice? *Statistics Education Research Journal, 15*(2). https://doi.org/10.52041/serj.v15i2.238

Pino-Fan, L. R., Font, V., Gordillo, W., Larios, V., & Breda, A. (2017). Analysis of the meanings of the antiderivative used by students of the first engineering courses. *International Journal of Science* *and Mathematics Education*, *16*, 1091-1113. https://doi.org/10.1007/s10763-017-9826-2

Pino-Fan, L. R., Godino, J. D., & Font, V. (2011). Faceta epistémica del conocimiento didáctico-matemático sobre la derivada [Epistemic facet of didactic-mathematical knowledge about the derivative]. *Educação Matemática Pesquisa, 13*(1), 141-178.

Pino-Fan, L. R., Godino, J. D., & Font, V. (2018). Assessing key epistemic features of didactic-mathematical knowledge of prospective teachers: the case of the derivative. *Journal of Mathematics Teacher Education, 21*(1), 63-94. https://doi.org/10.1007/s10857-016-9349-8

Pino-Fan, L. R., Guzmán, I., Font, V., & Duval, R. (2017). Analysis of the underlying cognitive activity in the resolution of a task on derivability of the absolute-value function: Two theoretical perspectives. *PNA, 11*(2), 97-124. https://doi.org/10.30827/pna.v11i2.6076

Presmeg, N. (2014) Semiotics in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education*. Springer. https://doi.org/10.1007/978-3-030-15789-0_137

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal, 5*(2), 46-68. https://doi.org/10.52041/serj.v5i2.500

Rossman, A. J. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal, 7*(2), 5-19. https://doi.org/10.52041/serj.v7i2.467

Rossman, A. J., & Chance, B. L. (2014). Using simulation-based inference for learning introductory statistics. *Wiley Interdisciplinary Reviews: Computational Statistics, 6*(4), 211-221. https://doi.org/10.1002/wics.1302

Sánchez Acevedo, N., & Ruiz Hernández, B. (2022). Análisis de las actividades que proponen dos libros de texto de educación primaria. Un acercamiento comparativo desde la perspectiva de la inferencia informal [An analysis of the activities proposed by two primary education textbooks. A comparative approach from the perspective of informal inference]. *Revista de Estudios y Experiencias en Educación* [*Magazine of Studies and Experiences in Education*], *21*(46), 76-101. https://doi.org/10.21703/0718-5162.v21.n46.2022.004

Sotos, A. E. C., 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. https://doi.org/10.1016/j.edurev.2007.04.001

Stohl Lee, H., Angotti, R. L., & Tarr, J. E. (2010). Making comparisons between observed data and expected outcomes: Students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal, 9*(1), 68-96. https://doi.org/10.52041/serj.v9i1.388

Trumpower, D. L. (2013). Formative use of intuitive analysis of variance. *Mathematical Thinking and Learning, 15*(4), 291-313. https://doi.org/10.1080/10986065.2013.830952

Trumpower, D. L. (2015). Aspects of first year statistics students' reasoning when performing intuitive analysis of variance: Effects of within-and between-group variability. *Educational Studies in Mathematics, 88*(1), 115-136. https://doi.org/10.1007/s10649-014-9574-y

Vallecillos, A (1994). *Estudio teórico experimental de errores y concepciones sobre el contraste de hipótesis en estudiantes universitarios* [*Theoretical experimental study of errors and conceptions on the contrast of hypotheses in university students*] [Doctoral dissertation, Universidad de Granada].

Vera, O. D., & Díaz, C. (2013). Dificultades de estudiantes de psicología en relación al contraste de hipótesis [Difficulties of psychology students in relation to hypothesis testing]. In J. M. Contreras, G. R. Cañadas, M. M. Gea, & P. Arteaga (Eds.), *Actas de las Jornadas Virtuales en Didáctica de la Estadística, Probabilidad y Combinatoria* [*Minutes of the Virtual Conference on Didactics of Statistics, Probability and Combinatorics*] (pp. 197-203). Universidad de Granada.

Vera, O. D., Díaz, C., & Batenero, C. (2011). Dificultades en la formulación de hipótesis estadísticas por estudiantes de psicología [Difficulties in the formulation of statistical hypotheses by students of psychology]. *UNIÓN. Revista Iberoamericana de Educación Matemática* [*UNION. Ibero-American Journal of Mathematics Education*], *27*, 41-61.

Vidal-Szabó, P., Kuzniak, A., Estrella, S., & Montoya, E. (2020). Análisis cualitativo de un aprendizaje estadístico temprano con la mirada de los espacios de trabajo matemático orientado por el ciclo investigativo [Qualitative analysis of early statistical learning with a view to mathematical workspaces guided by the research cycle]. *Educación Matemática* [*Mathematics Education*], *32*(2), 217-246. https://doi.org/10.24844/EM3202.09

Weinberg, A., Wiesner, E., & Pfaff, T. J. (2010). Using informal inferential reasoning to develop formal concepts: Analyzing an activity. *Journal of Statistics Education, 18*(2). https://doi.org/10.1080/10691898.2010.11889494

Wild, C. J., Utts, J. M., & Horton, N. J. (2018). What is statistics? In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 5-36). Springer. https://doi.org/10.1007/978-3-319-66195-7_1

Wittgenstein, L. (1953). *Philosophical investigations*. The MacMillan Company.

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40-58. https://doi.org/10.52041/serj.v7i2.469

https://www.ejmste.com