



Revemop

ISSN: 2596-0245

revemop@ufop.edu.br

Universidade Federal de Ouro Preto
Brasil

Lugo-Armenta, Jesús Guadalupe; Pino-Fan, Luis R; Hernández, Blanca R. Ruiz
Chi-square Reference Meanings: a Historical-epistemological Overview
Revemop, vol. 3, 2021, January-December
Universidade Federal de Ouro Preto
Brasil

DOI: <https://doi.org/10.33532/revemop.e202108>

- Complete issue
- More information about this article
- Journal's webpage in redalyc.org



Chi-square Reference Meanings: a Historical-epistemological Overview

Jesús Guadalupe Lugo-Armenta Doutorando em Educação Matemática pela Universidad de Los Lagos, *campus Osorno*. Osorno, Los Lagos, Chile.

 <http://orcid.org/0000-0001-6679-5115>
 jesus.lugo@ulagos.cl

Luis R. Pino-Fan Doutor em Didáctica de la Matemática pela Universidad de Granada. Professor da Universidad de Los Lagos, Osorno, Los Lagos, Chile.

 <http://orcid.org/0000-0003-4060-7408>
 luis.pino@ulagos.cl

Blanca R. Ruiz Hernández Doutora em Didáctica de la Matemática pela Universidad de Granada. Professora do Tecnológico de Monterrey, Monterrey, Nuevo León, México.

 <http://orcid.org/0000-0003-0157-3866>
 bruiz@tec.mx

Recebido em 10/05/2021

Aceito em 07/06/2021

Publicado em 21/06/2021

Abstract: The present article shows a historical-epistemological study on the Chi-square statistic. In which theoretical-methodological notions from the Onto-Semiotic Approach (OSA) of mathematical cognition and instruction were used to identify four problems that have been key to the evolution of the Chi-square statistic: the Goodness-of-fit-test, the test of independence, the test of homogeneity and distribution. Furthermore, various meanings of the Chi-square statistic were recognized in the mathematical-statistical practices that are used to solve each of those problems. These meanings could help to establish epistemic criteria that allow, on the one hand, to propose progressive levels of inferential reasoning for the statistic (from informal to formal); and on the other hand, to design tasks oriented to promote the understanding of the diverse meanings of the Chi-square.

Keywords: Chi-square. History and epistemology. Inferential reasoning. Meanings. Statistics education.

Significados de Referência da Estatística Qui-quadrado: um Olhar Histórico-epistemológico

Resumo: Este artigo apresenta um estudo histórico-epistemológico sobre a estatística Qui-quadrado. Para tanto, são utilizadas algumas noções teórico-metodológicas da Abordagem Onto-Semiótica (EOS) do conhecimento e do ensino matemático, que nos permitiram identificar quatro problemas que têm sido fundamentais para a evolução da estatística Qui-quadrado: teste de adequação, teste de independência, teste de homogeneidade e distribuição. Além disso, nas práticas matemático-estatísticas realizadas para resolver cada um destes problemas, foram identificados vários significados da estatística Qui-quadrado, o que permitirá estabelecer critérios epistemológicos que permitem, por um lado, propor níveis progressivos (do informal ao formal) do raciocínio inferencial para a referida estatística; e por outro lado, desenhar tarefas que visem promover a compreensão dos diversos significados do Qui-quadrado.

Palavras-chave: Qui-quadrado. História e epistemologia. Raciocínio inferencial. Significados. Educação estatística.

Significados de Referencia del Estadístico Chi-cuadrada: una Mirada Histórico-epistemológica

Resumen: El presente artículo muestra un estudio de tipo histórico-epistemológico sobre el estadístico Chi-cuadrado. Para ello, se utilizan algunas nociones teórico-metodológicas del Enfoque Onto-Semiótico (EOS) del conocimiento y la instrucción matemática, las cuales nos permitieron identificar cuatro problemáticas que han resultado clave para la evolución del estadístico Chi-cuadrada: prueba de bondad de ajuste, prueba de independencia, prueba de homogeneidad y distribución. Además, en las prácticas matemáticas-estadísticas llevadas a cabo para resolver cada una de estas problemáticas, se identificaron diversos significados del estadístico Chi-cuadrada, los cuales permitirán establecer criterios epistemológicos que permitan, por un lado, proponer niveles progresivos (de lo informal a lo formal) del razonamiento inferencial para dicho estadístico; y por otro, diseñar tareas orientadas a promover la comprensión de los diversos significados de la Chi-cuadrada.

Palabras clave: Chi-cuadrada. Historia y epistemología. Razonamiento inferencial. Significados. Educación estadística.

1 Background

Currently, Statistics and specifically Inferential Statistics, has taken a crucial role in professional, as well as, in the daily life of a significant number of people, as the world in which they develop is rapidly changing, originating information and data that quickly expands and varies every day. However, several investigations (e.g., SALDANHA; THOMPSON, 2002; BAKKER; GRAVEMEIJER, 2004; ROSSMAN, 2008; REABURN, 2014; HOEKSTRA, 2015), have reported difficulties for students of different educational levels in comprehending and connecting notions considered as essential for Inferential Reasoning, such as Variation, distribution, sampling, sampling variability, sampling distribution, p-value, level of significance, construction of hypothesis, critical values in statistical distribution, and statistic-parameter.

In this sense, new proposals on how to work Inferential Reasoning have emerged from an informal and formal perspective. On the one hand, the perspective of Informal Inferential Reasoning (IIR) aims at integrating and giving meaning to statistical concepts generating an early contact with Inferential Statistics for students (e.g., ZIEFFLER et al., 2008; MAKAR; RUBIN, 2009; DOERR et al., 2017). On the other hand, from the perspective of Formal Inferential Reasoning (FIR), research has been conducted (e.g., TARLOW, 2016; RIEMER; SEEBACH, 2014; ROCHOWICZ, 2010) on student's comprehension of the formal methods of statistical inference; for example, the reflection about the logic of hypothesis testing, the different moments of decision-making, and p-value.

The teaching of statistics can address the challenge of introducing the notions coherently from an IIR perspective, starting from the intuitive aspects. Nevertheless, Batanero (2013) makes

a call for reflection on the exact level of formalization required to teach the statistical notions. Some investigations (e.g., JACOB; DOERR, 2014; PFANNKUCH et al., 2015; MAKAR; RUBIN, 2018), indicate the need to promote the FIR progressively in students. In order to promote FIR progressively, it is primarily necessary to comprehend how the statistical notions emerge from a range of mathematical practices that helped to solve problems from different scientific fields, which allows for the identification of diverse meanings for the same notion. This article focuses on the study of the Chi-square statistic considering its importance in the construction of a methodology for hypothesis tests and its current relevance in Statistics Education.

Thus, this article aims at proposing a reconstruction of the holistic meaning attributed to the Chi-square statistic; by distinguishing both the different critical problems for its emergence and development and the characterization of the diverse meanings that it has had throughout history. We consider that the characterization of the meanings for this statistic helps to retrieve its mathematical richness and allows rescuing elements from history to propose progressive levels (from informal to formal) of Inferential Reasoning of the Chi-square statistic. Which, in turn, will enable the design of activities for the gradual comprehension of the statistic.

2 Theoretical and methodological framework

To carry out this study, we used some theoretical-methodological tools from the Onto-Semiotic Approach (OSA) of mathematical cognition and instruction (GODINO et al., 2007). The OSA allows to make a detailed analysis of the Chi-square statistic meanings through a profound description of the problems, the mathematical practices developed to solve such problems, and the mathematical objects that intervene in those practices. The notion of system of practices plays a crucial role in this theoretical approach. It refers to "any action or manifestation (linguistic or otherwise) carried out by somebody to solve mathematical problems, to communicate the solution to other people, to validate and generalize that solution to other contexts and problems" (GODINO; BATANERO, 1994, p. 334). The mathematical practices can be personal or shared by a group within an institution (institutional practices). Godino and Batanero (1994), define the institutional practices as "The institutional system of practices, associated to a field of problems, it is constituted by the practices considered as significative to solve a field of problems C and shared in the heart of an institution I" (p. 337).

In the mathematical practices adopted to solve a determined field of problems, mathematical objects intervene and emerge. The mathematical objects can be ostensive (like

symbols, graphs) and non-ostensive (as concepts, propositions). They can emerge from institutional systems of practices (institutional objects), or personal systems of practices (personal objects). The OSA framework also proposes types of primary mathematical objects intervening in the system of practices (GODINO et al., 2007): linguistic elements, situations/problems, concepts/definitions, propositions/ properties, procedures, and arguments. The primary mathematical objects are interrelated, forming frameworks called configurations, which can be epistemic if the frameworks of objects are an institutional practice, or cognitive if those frameworks configure a personal practice. The notion of configuration "responds to the need to identify the types of objects and processes that intervene and emerge in the mathematical practices used to solve the situations-problems" (GODINO et al., 2019, p. 39).

In this sense, in OSA, the meaning of mathematical objects is conceived from a pragmatic-anthropological perspective, which considers the relativity of the context in which these are used. In other words, the meaning of a mathematical object can be defined as the system of operative and discursive practices that a person (or an institution) develops in order to solve a particular type of situations-problems in which such object intervenes (GODINO; BATANERO, 1994). Thus, the meaning of a mathematical object can also be considered from two perspectives, institutional and personal. The notion of institutional meaning allows studying the practices where the mathematical objects emerge, its historical evolution, and it also addresses the contexts. This type of study is called historical-epistemological, and through them, it is possible to determine the holistic meaning, which comprises various partial meanings of a mathematical object (PINO-FAN et al., 2011), which in turn have an epistemic configuration associated with them.

The notion of onto-semiotic configuration has been employed in different investigations to characterize the holistic meaning of diverse mathematical notions, mainly of calculus (e.g., GODINO et al., 2011; PINO-FAN et al., 2018). It was used in the present study, as it allowed us to locate the problems that were key to the development of the Chi-square statistic and how they were solved. Those problems and the respective mathematical practices used to solve them are associated with an epistemic configuration, which helps us to determine a partial meaning of the statistic mentioned.

This study employs a qualitative methodology (COHEN et al., 2011) of historical-documentary type, in which primary sources, original texts, secondary sources, mathematics, and statistics history books, were revised.

3 Historical evolution of the Chi-square statistic

We identified three main problem areas in the development and evolution of the chi-square statistic (χ^2): the goodness-of-fit-test, independence, and homogeneity. Additionally, the importance that the χ^2 distribution has in these tests is considered as the asymptotic distribution of the χ^2 statistic.

3.1 The beginnings of the goodness-of-fit-test

In the eighties of the XIX century, Francis Galton, Ysidro Edgeworth, and Karl Pearson initiated the works that provided the main contributions for the construction of an empirical and conceptual methodology for statistics. In 1875, Galton presented the method of intercomparison, through which he organized the data in increasing order and graphed the data values versus the ranges (quartiles). He called this graphic ogive. In the ogive m at $\frac{1}{2}$ represents the mean value of the series (the median), p at $\frac{1}{4}$ and q at $\frac{3}{4}$ provide data for estimating the divergence taken in connection with m ; accordingly, $q - m$ is the divergence or probable error (of that portion of the series that exceeds the mean), while $m - p$ corresponds to other portion of the series. According to Galton (1875), when the series is symmetrical $q - m = m - p$.

Galton (1875) worked in the opposite direction of the law of frequency of error, and stated that, "since such and such magnitudes occur with such and such degrees of frequency; therefore the differences between them and the mean value are so and so, as expressed in units of probable error" (p. 38). In 1885, Galton worked on the graphic method, which consisted of two parts. In the first, he resumed the method of intercomparison, except that he was no longer working uniquely with quartiles in the base of the ogive, but used the percentiles and considered the quartile two or the median as 0 in the sense of the works developed by Quetelet, where the other elements of the set are considered deviations of the representative to ± 50 . In the second part, he indicated that to apply this test it was necessary to change the line of reference from which the ordinates were originated, parting the curve of distribution (the ogive) in two parts, "and the lowermost reversed" (GALTON, 1885 p. 262). He also provided a table of the normal curve of distribution of error and commented that:

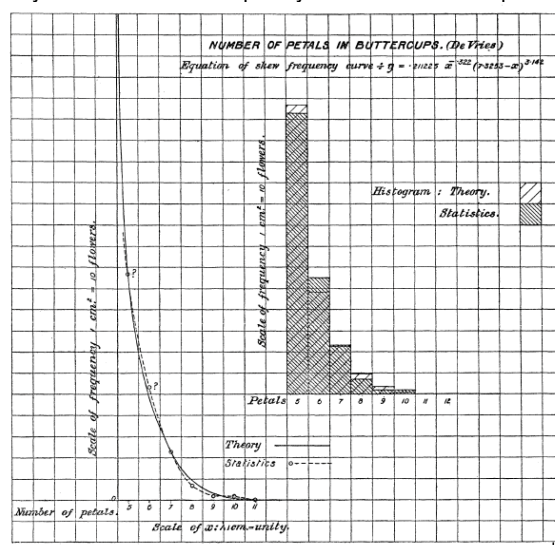
In order to bring the observed values into a form suitable for comparison with this table, we must begin by measuring the observed deviates at $\pm 10^\circ, 20^\circ, 25^\circ, 30^\circ, 40^\circ$, and 45° [...], if the series is "normal," the values so obtained will be identical with those in column B, and if it is approximately normal, they will correspond approximately. (GALTON, 1885, p. 265)

In the subsequent years, the graphic method of Galton was widely used, in his works and others'. For example, in 1890, Weldon adjusted the data about the carapace lengths of the Plymouth shore crab to the normal curve following that method as it was then appropriate for the data. However, in later work in 1892, when he was analyzing the double-humped curve carapace in female shore crabs of Naples, noticed that Galton's graphic method was not suitable enough for this not-normally-distributed data. Following Hald (2007), Pearson helped Weldon with the analysis of zoological data for studies on evolution. From this, Pearson felt the need for a series of continuous distributions to describe the biological phenomena that he was studying, and in 1895 he published a system or family of continuous probability distributions (seven types of frequency curves). In this distribution system, he used the same measures of skewness and kurtosis as Thiele, but he expressed them in terms of moments.

In 1894, in an early consideration of the goodness-of-fit-test with the χ^2 statistic, Pearson "demonstrated how to find $\sum s/y$, where s equalled the difference between the observation polygon and the theoretically expected curve and y its corresponding ordinate" (MAGNELLO, 2005, p. 727). Measuring the relation of the whole area between the curve and the polygon, where all the positive values equaled W and A was the area under the curve; obtaining the following expression $W/A = (\sum \text{errors of fit})/(\sum \text{ordinates}) = \sum s/\sum y$, where s are the differences between the observed frequencies and the expected frequencies and y its correspondent ordinate.

The following data was presented for Pearson in 1895 and corresponds to the number of petals in a garden flowers (observed frequency) and calculate the areas and results the theory frequency. Further, the comparison between theory and observation is represented by curve and histogram in figure 1.

Figure 1: Theory and observation frequency curve of number of petals in buttercups



Source: Pearson (1895, plate 15)

Pearson tested his system of continuous probability distribution with a variety of data from meteorology, anthropometry, zoology, botany, economy, demography, and mortality, showing that his distributions adjusted to the data (HALD, 2007). However, Pearson continued working on an objective or reasonable measure of the goodness-of-fit, a problem that he tried to solve in 1900 with his proposal of the goodness-of-fit-test through the χ^2 statistic (as hypothesis testing). Made possible to observe to what extent an observed dataset is adjusted to pre-established theoretical distribution, through the contrast of the observed frequencies and the expected frequencies. The χ^2 statistic follows a distribution of the probability of the same name and is given by Pearson type III distribution. Pearson (1900), indicated that the objective was "to investigate a criterion of the probability on any theory of an observed system of errors, and to apply it to the determination of goodness of fit in the case of frequency curves" (p. 157). For this, Pearson starts from the expression for the normal density function of an n vector, of variables x , with mean $\mu = 0$ and dispersion matrix V . According to Barnard (1992), the current notation is: $K \exp^{-\frac{1}{2}(x - \mu)'V^{-1}(x - \mu)}$, where $'$ denotes transpose, and K is determined by the condition that density integrates to 1. Pearson indicated that: $\chi^2 = x'V^{-1}x$; he also established that χ^2 equals a constant and represented the equation of a generalized ellipsoid on the sample space and that the values that χ^2 takes must be given to the space range from 0 to ∞ , that is to say, $P(x > x_0)$ is calculated. If the ellipsoid becomes a sphere, then X 's would refer to coordinates, and the chances of a system of errors with high frequencies would be denoted by χ . Pearson also made a transformation to generalized polar coordinates, where χ could be treated as one of the lines that

diverge from a common center and obtained: $P = \frac{\int_{\chi}^{\infty} e^{-\frac{\chi^2}{2}} \chi^{n-1} d\chi}{\int_0^{\infty} e^{-\frac{\chi^2}{2}} \chi^{n-1} d\chi}$. In this regard, Pearson indicated

that:

This is the measure of the probability of a complex system of n errors occurring with a frequency as great or greater than that of the observed system [...] and then an evaluation of $[P]$ gives us what appears to be a simple fairly reasonable criterion of the probability of such an error occurring on a random selection being made. (PEARSON, 1900, p. 158)

Then he applied the results to the problem of the fit of an observed to a theoretical frequency distribution. If the data are grouped in $n + 1$, then the observed frequencies of the groups would be: $m'_1, m'_2, m'_3, \dots, m'_n, m'_{n+1}$; while the theoretical frequencies supposedly known previously are: $m_1, m_2, m_3, \dots, m_n, m_{n+1}$. As $e = m' - m$ give the error, then

$e_1, e_2, e_3, \dots, e_n, e_{n+1} = 0$ if the observed frequency corresponds to the theoretical. From this and the initial consideration of χ^2 , it was possible to obtain: $\chi^2 = S \left(\frac{e^2}{m} \right)$.

3.2 The beginnings of the test of independence

George Udny Yule was looking for a way to measure the association of attributes statistically. For example, if there was an association (in the sense of correlation, but for discrete variables) between medical conditions like deafness, blindness, and intellectual disability or cases of mortality from diseases and the administration of new antitoxins.

To approach this problem, Yule (1900) starts from the theory of statistical correlation and provided examples of correlation of two types, one corresponds to continuous variables (e.g., measurement on portions of the body) and the other to discrete variables (e.g., number of children in a family). To justify and develop his work of association, he began from theorems of the set theory, signaling that "two qualities or attributes, A and B, are defined to be independent if the chance of finding them together is the product of the chances of finding either of them separately, i.e., if $\frac{(AB)}{U} = \frac{(A)}{(U)} \cdot \frac{(B)}{(U)}$ or $(AB)(U) = (A)(B)$ " (YULE, 1900, p. 270). Furthermore, he indicated that this was the only legitimate test of independence or association (understanding association as dependence). Yule (1900, p. 272) established that $Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$. Regarding Q, Yule (1900, p. 271) indicated that he intended it to "be a measure of the approach of association towards complete independence on the one hand and complete association on the other."

In 1904 Pearson published an article on contingency and its relation to association and normal correlation. In this paper, he resumed both, his works on the χ^2 statistic from 1900 and Yule's work on association developed the same year. In the same work, Pearson (1904) also indicated that it is possible to classify the attributes in different groups in a table, thus constituted by s columns and t rows, with the universe (frequency of the population) distributed into sub-groups corresponding to the $s \times t$ compartments. Pearson recognized the problem he was facing when there are high numbers of attributes, as in the classification of human eye color into eight types and the correlation of these with six classes for hair color. To address this type of problem, he introduces the concept of contingency. Furthermore, to make connections with the notions of association and normal correlation, he proceeded from probabilistic independence, indicating that "if p be the

probability of any event, and q the probability of a second event, then the two events are said to be independent, if the probability of the combined event be $p \times q$ " (PEARSON, 1904, p. 5). According to Pearson, when we measure contingency, we measure the deviations of the observed results of the independent probability, thus, if $m_1, m_2, \dots, m_n$ correspond to the theoretical system v_{uv} and $m'_1, m'_2, \dots, m'_n$ belonging to the observed system n_{uv} , then $\chi^2 = S\{(n_{uv} - v_{uv})^2 / v_{uv}\}$. In this case, P is a measurement of how far the observed system is, thus determining the compatibility or not with the bases of probabilistic independence. To calculate probability, he established an $n' = (x - 1)(\lambda - 1) / N - \phi^2$, where x is the number of rows and λ the number of columns, so, when P is large the chances are in favor of the system originating from independent probability. When P is small there is association between the attributes. Thus, Pearson recognized $1 - P$ as an appropriate measure of the contingency, which he called the contingency grade.

Clearly the greater the contingency, the greater must be the amount of association or of correlation between the two attributes, for such association or correlation is solely a measure from another standpoint of the degree of deviation from independence occurrence. (PEARSON, 1904, p. 6)

Pearson also formulated a function related to the χ^2 statistic and called it the mean square contingency $\phi^2 = \frac{\chi^2}{N}$, and termed mean contingency to the summation of all the positive contingencies $\psi = \sum \frac{n_{uv} - v_{uv}}{N}$. Furthermore, he pointed out that "any functions of either, ϕ^2 or ψ would serve to measure the contingency" (PEARSON, 1904, p. 7). He also expressed ϕ^2 and ψ in terms of association for two attributes established by Yule (1900), in the following way: $\phi^2 = \frac{(ab - cd)^2}{(a + d)(c + b)(a + c)(d + b)}$ and $\psi = \frac{2(ab - cd)}{N^2}$. Pearson showed how ϕ^2 and ψ were closely connected with Yule's notion of association.

3.3 Test of homogeneity

Pearson (1911), took as a base the problem that triggered the treatment of goodness-of-fit of 1900, however, indicated that the treatment would be different as, "we have two samples, and a priori they may be of the same population or of different populations; we desire to find out what is the probability that they are random samples of the same population. This population is one, however, of which we have no a priori experience" (p. 250). He indicated examples of the type of problems in which it would be desirable to have such information about the samples. Like with the

given records of the number of rooms in houses where some of the inhabitants had had cancer or tuberculosis. He added that the number of cases of each disease may be very different and that the frequency distribution of the number of rooms in the given district may be unknown.

From this problem, Pearson posed the following question: What is the chance that there is a significant difference in the houses with cases of tuberculosis and the ones with cancer cases? He started from the premise that the population, from which the two samples are drawn, can be given by the class $\mu_1, \mu_2, \mu_3, \mu_4 \dots \mu_p, \mu_q \dots \mu_g$, being M the total population and where the samples would be given by the frequencies in the same classes. He also supposed that the two samples were independent, and thus, there would be no correlation between any deviation in any frequency of the first and second row. Pearson (1911) indicated that the statistic has the form:

$$\chi^2 = S_1^g \left\{ \frac{NN' \left(\frac{f_p}{N} - \frac{f_p'}{N'} \right)^2}{f_p + f_p'} \right\}.$$

In 1930, George Snedecor presented a statistical technique that he called experimental, the statistical test of homogeneity. He indicated that the test was applicable to the type of data known as homograde statistics or the statistics of attributes and that it worked with this type of data when the elements or individuals in a sample can be classified in categories. While in 1933, Snedecor and Irwin in the article 'On the Chi-square test for homogeneity' presented a proposal for the test of homogeneity with the χ^2 statistic. These authors used the expression $\chi^2 = \frac{100 (\sum sp - \bar{p} \sum s)}{\bar{p}q}$, to calculate the statistical value; and to obtain the probability, they used the tables of the distribution χ^2 with $n' = n + 1$, where n was the degrees of freedom, which were obtained from $n = (c - 1)(r - 1)$. Snedecor and Irwin (1933), indicated that if the data can be divided into subsets, it could be possible to determine the probability that the series of subsamples may have been extracted from a homogeneous population. This form of expressing probability is different from interpreting probability in favor or against an event and as the mean probability of the whole set.

3.4 The refinements of the goodness-of-fit, independence and homogeneity tests

Greenwood and Yule (1915), observed a discrepancy when they compared the results obtained through the test with the χ^2 statistic and the test through the difference of proportions and its probable error, and added that such discrepancy "is surprising and is not due to the neglect of the correlation in errors between the sub-group frequencies" (p. 118).

Fisher (1922), starts from the previous difficulty highlighted by Greenwood and Yule and indicates that "when we recognize that we should take $n' = 2$, the difficulty disappears" (p. 90). While for Greenwood and Yule (1915), $p = \frac{a}{a+b}$ and $p' = \frac{c}{c+d}$, for Fisher (1922) the standard error is $p = \sqrt{\frac{(a+c)(b+d)}{(a+b+c+d)^2(a+b)}}$ and $p' = \sqrt{\frac{(a+c)(b+d)}{(a+b+c+d)^2(c+d)}}$. So, if $x = p' - p$, then we obtain: $\frac{x^2}{\sigma_x^2} = \frac{(bc-ad)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)} = \chi^2$ when working with 2×2 tables, while the expression $\chi^2 = \sum \frac{((m+x)-m)^2}{m}$ was used with $c \times r$ contingency tables. This last expression about the χ^2 statistic was used by Fisher in the three tests, goodness-of-fit, independence, and homogeneity.

Fisher (1922) indicated that, in general, the χ^2 test introduced by Pearson is adequate if the number of observations is large. Fisher used Elderton's goodness-of-fit tables from 1914 but introduced a variation in the form of calculating n' . He indicated that for a contingency table of c columns and r rows $n' = (c-1)(r-1) + 1$. Although Yule, Greenwood, Pearson, and Elderton carried out works and approaches on n' for the χ^2 distribution, Fisher is credited with introducing the number of degrees of freedom, and defines them, on the one hand, for the goodness-of-fit-test as the number of rows minus the number of independent linear restrictions (r) in the frequencies (i.e., $gl = n - r$); on the other hand, for the test of independence for the contingency tables with c columns and r rows $n' = (c-1)(r-1) + 1$.

In 1925, Fisher provided the means to apply statistical tests with precision, defining the contrasts of signification where there was only one hypothesis -the null hypothesis-, which specified the numerical value of the parameter. The critical ideas of Fisher's experiments, developed in the agriculture field, were the randomization and the significance. In 1934, Fisher provided a table for the χ^2 distribution, where it was possible to find the value of the χ^2 statistic according to the level of significance and degrees of freedom of the test, with the value of the χ^2 statistic found in the table, the zones of rejection and acceptance are established.

Wijayatunga (2016) performs an interpretation of the chi-square statistic, when it is used in the independence test for two discrete random variables, with conditional probability, arguing that this test uses a certain measure of dependence. He indicates that X take values $i = 1, \dots, \alpha$ and Y take values $j = 1, \dots, \beta$. The joint probability of $X = i$ and $Y = j$ as p_{ij} , the marginal probability of $X = i$ and $Y = i$ as $p_{i.}$ and $p_{.j}$ respectively. And, the conditional probability of $X = i$ given by

$$Y = j \quad \text{is} \quad p_{i|j} = p_{ij}/p_{.j}. \quad \text{Then,} \quad \chi^2 = \sum_{i,j} n \frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}} = n \left\{ \sum_{i,j} p_{ij} \frac{p_{ij}-p_{i.}p_{.j}}{p_{i.}p_{.j}} \right\} =$$

$n \left\{ \sum_{i,j} p_{ij} \frac{p_{ij} - p_{i.}}{p_{i.}} \right\} = n \left\{ \sum_{i,j} p_{ij} \frac{p_{.j} - p_{.j}}{p_{.j}} \right\} = nE\{A\}$, where A is a random variable with probability p_{ij} and E denotes the expectation. Also, $\frac{p_{ij} - p_{i.}}{p_{i.}}$ may refer to the degree of dependence between X and Y . While $E\{A\}$ can be interpreted as a measure of degree of dependence between X and Y .

Regarding the homogeneity and independence tests with the χ^2 statistic, Fisher (1934) pointed out that both tests are mathematically identical, that the " χ^2 index of dispersion would then be equivalent to the χ^2 obtained from the contingency table" (p. 94) and, also, that the homogeneity test might be applied to samples of equal or different sizes. Additionally, Yates (1934), proposed a continuity correction factor for the χ^2 statistic, in the context of the test of independence. With the correction for continuity, it was possible to use the test of independence with small numbers. Yates (1934, p. 217) indicates that "the accuracy of this approximation depends on the numbers in the various cells, and in practice, it has been customary to regard χ^2 as sufficiently accurate if no cell has an expectancy of less than 5". Thus, he focused his work on the applicability of the tests to contingency tables with low expectations.

According to Yates (1934), the discrepancies are because χ is a continuous distribution, while the distribution it is attempting to approximate is discontinuous. When the values of χ^2 are calculated for deviations, half a unit less than the right deviations, it is denominated correction for continuity and the resultant value of χ is χ' and with a $P = (\chi')$. Namely, this correction for continuity consists in subtracting 0.5 to the positive deviations and adding 0.5 to the negative deviations. He starts from $\chi^2 = \sum \frac{(a \times d - b \times c)^2 N}{n \times n' (N - n) (N - n')}$ to establish that the statistic with correction is

$$\chi_c^2 = \sum \frac{\left(\left(a - \frac{1}{2} \right) \left(d - \frac{1}{2} \right) - \left(b + \frac{1}{2} \right) \left(c + \frac{1}{2} \right) \right)^2 N}{n \times n' (N - n) (N - n')}, \text{ when working with } 2 \times 2 \text{ contingency tables.}$$

Yates (1934), examined the discrepancies with the χ^2 statistic and its associated probability, after the correction for continuity in contingency tables that involve small numbers with one or more degrees of freedom. The application of the continuity correction factor was expanded to the goodness-of-fit and homogeneity tests.

4 Meanings of the Chi-square statistic

In the present section, the onto-semiotic configurations identified with the historical analysis of the χ^2 statistic will be described. The segment initially addresses the three main problem areas

that allowed the emergence, development, and generalization of this mathematical object. That is to say, the goodness-of-fit-test, the test of independence, and the test of homogeneity. These problem areas are closely related to the problem: χ^2 distribution. The holistic meaning of the χ^2 statistic is comprised of twelve partial meanings. To determine each partial meaning, an onto-semiotic configuration has been described from the mathematical practice used to solve the previously mentioned problem areas.

4.1 Problem area 1: goodness-of-fit-test

In the historical development of the goodness-of-fit test with the χ^2 statistic, four partial meanings were identified. The first one corresponds to an intuitive goodness-of-fit-test since the χ^2 statistic was still not explicitly used in the practices.

4.1.1 Partial meaning 1 (PM1): Galton's graphic method

Galton was looking for a way to establish if it was appropriate to work an observed dataset under the conditions of the normal distribution. To solve this type of problem, in 1875, he developed the method of intercomparison, and in 1885 he presented the graphic method. To explain the graphic method, Galton worked with a dataset observed in 775 women. The data related to the height of female adults aged 23 to 50. To prove that the data followed a normal distribution, the graphic method was applied.

We can highlight from Galton's graphic method, the use of some linguistic elements like natural language in the formulation and interpretation of results, tabular representation to show the observed frequencies of height (sitting) of the 775 females, and the use of graphic representation (ogive) under the method of intercomparison and to show the deviates distribution regarding the 50° percentile. Furthermore, concepts/definitions were identified in the mathematical practice including the introduction of the ogive, quartiles, percentiles, modulus, mean error, and, notably, the probable error as a measure of the variability of the observed series; as well as the deviates which considered the new ordinates of the graph and represented the differences between each observed value and the m value of all the observed values (also denominated divergencies or errors).

It was also possible to identify properties/propositions as the law of error, normal distribution, m in $\frac{1}{2}$ or at 0° , p in $\frac{1}{4}$ or in -25° , q in $\frac{3}{4}$ or at 25° , probable error, and it is highlighted that if the series is symmetric, then $q - m = m - p$. The procedures of this configuration are the

graphing of the ogive of cumulative frequencies and the graphing of the distribution of deviations between each observed value and m . From these observed properties that helped to build the aforementioned graphics, the arguments are focused on establishing if the observed datasets of 775 females follow a normal distribution.

4.1.2 Partial meaning 2 (PM2): Goodness-of-fit test with Pearson's χ^2 statistic

In 1900, Pearson tried to solve the problem of an objective or reasonable measurement of the goodness-of-fit with his proposal of the goodness-of-fit-test through the χ^2 statistic. To illustrate the usefulness and the way of applying the goodness-of-fit-test, Pearson presented some prototype-problems. For example, he resumed the data about Merriman's 1000 shots. Merriman participated in the Spanish Civil War, hence his interest for analyzing deviations of 1000 shots during military training, he used the Method of Least Squares and indicated that the data derived from a probable system of deviations from the normal curve. However, Pearson (1900), resumes this problem to prove if the data of the 1000 shots follow a normal distribution, as suggested by Professor Merriman, through the χ^2 goodness-of-fit-test.

The linguistic elements used to develop the goodness-of-fit-test are natural language, in the formulation and interpretation of results, tabular representation to show the frequency distributions (theoretical and observed), as well as for calculating error and squared error between the theoretical frequency. Additionally, symbolic representation is used in the table and in the case of calculating to obtain the probability through the formula. The theoretical and observed frequencies shown through the tabular representation constitute some concepts/definitions of this configuration, like the experiments, experiments tests, probability and the χ^2 statistic, which is a measure of the divergence between the theoretical frequencies distribution and the observed ones.

Some properties/propositions observed in the tabular representation were the error (i.e., the differences between the observed frequency (m') and the theoretical frequency (m) under the expression $e = m' - m$); the χ^2 probability distribution as an asymptotic distribution of the χ^2 statistic, the number of errors n' , the formulas for the calculation of probability under χ and n' , and the identified elements resumed in the PM1, as the law of error and normal distribution. The arguments are provided from the results of the test and in terms of the context of the problem, are focused on finding the probability of occurrence of the deviations system of the set of observed frequencies respecting the expected frequencies or theoretical of particular distribution (in this case of the normal distribution), for which it is necessary to calculate the χ^2 statistical value under the

following expression which results in a property/proposition $\chi^2 = \sum \frac{(m' - m)^2}{m}$. Some procedures used for this are the calculation of error and squared error between the expected frequency, which are also seen in the tabular representation.

4.1.3 Partial meaning 3 (PM3): Goodness-of-fit-test with Fisher's degrees of freedom

The problem that Fisher and other statisticians of that time addressed refers to the degrees of freedom for calculating the probability associated with the χ^2 statistical value in the χ^2 distribution, in the context of the goodness-of-fit and independence tests. In 1925 Fisher published some situations/problems about the goodness-of-fit-test application with the χ^2 statistic and with $n' = k + 1$, where k indicates the number of degrees of freedom. Notably, a problem where Fisher made a comparison with the expectation of Mendelian class frequencies in the generation of a hybrid (F1) of four classes in the ratio 9 : 3 : 3 : 1 is resumed. He introduces the null hypothesis or the researcher hypothesis: the factors segregate independently, and the four classes of descendants are equally viable.

To make the goodness-of-fit-test linguistic elements like natural language were used, in the same sense that in the PM2, tabular representation to show the frequency distributions (expected and observed) and the value of the statistic. Symbolic representation is used in the tables, in the formulation of the problem and conclusion. In the mathematical practice that Fisher extended to solve this problem, concepts/definitions of the PM2 were identified, e.g., experiment, test, observed frequency, expected frequency, probability. It was also possible to identify the introduction of the degrees of freedom, in this context, as the number of rows minus the number of independent linear restrictions in the frequencies; this concept was related to a fundamental property/proposition in this configuration, which is $k = n - r$. This property generated a change in n' , which is now known as the number of degrees of freedom plus one.

To solve the problem, procedures like calculating the value of the χ^2 statistic, establishing the number of degrees of freedom and n' to obtain the probability of the Elderton's tables of the χ^2 distribution were used. It is argued using the obtained probability like the one of occurrence having a dataset with a system of errors of such broad deviations of particular theoretical distribution; additionally, if prior the test, a limit was established as significant deviation (e.g., $P = 0.05$), the probability obtained from the tables was compared with such limit.

4.1.4 Partial meaning 4 (PM4): Goodness-of-fit-test with Yates' continuity correction factor

The problem that Yates addressed refers to the correction for continuity that the χ^2 statistic requires when working with small numbers. He indicated that the discrepancy becomes visible when working with small numbers because χ is a continuous distribution, while the distribution it is attempting to approximate is discrete. To illustrate Yates' method in a goodness-of-fit-test when working with frequencies below five, a Pearson (1900) situation/problem is resumed, which revolves around the number of petals in a flower denominated ranunculus, the expected distributions were calculated under the skewed curve $y = 0.211225x^{-0.322}(7.3253 - x)^{3.142}$.

Just as in the previous partial meanings it was possible to identify linguistic elements like natural language, tabular representation and the use of symbolic representation to indicate the value of the χ_c^2 statistic, degrees of freedom, and probability. Furthermore, concepts/definitions were resumed from PM2, such as experiment, test, observed frequency and expected frequency, probability, correction for continuity, and degrees of freedom from PM3. The correction for continuity factor was introduced as an adjustment made to the χ^2 statistic when working with small numbers, because it is intended to approximate a discontinuous distribution through a continuous distribution. So, this concept impacts the statistic directly generating the property/proposition $\chi_c^2 = \sum \frac{(|Observed\ frequency - Expected\ frequency| - 0.5)^2}{Expected\ frequency}$. Other properties/propositions of this configuration are in accordance with the established in other partial meanings, particularly, the law of error from PM1, the error and χ^2 probability distribution from PM2, and the degrees of freedom from PM3.

The procedures used to carry out the χ_c^2 goodness-of-fit-test with correction for continuity factor are: (1) calculating the value of the χ_c^2 statistic; (2) obtaining n' (in the sense of PM3); and (3) looking for the probability value in the χ^2 distribution table (with χ_c^2 and n'). The arguments developed in accordance with such probability value identified in the tables, indicating if the probability of occurrence having a dataset corresponds to particular theoretical distribution, for example, if having $\chi_c^2 = 2.587262$ and $P = .858576$, it can be said that the skewed curve $y = 0.211225x^{-0.322}(7.3253 - x)^{3.142}$ proposed by Pearson, is a proper adjustment for the distribution of the ranunculus petals.

4.2 Problem area 2: test of independence

Four partial meanings were identified in the evolution of the test of independence with the χ^2 statistic. We will refer to them continuing with the numbering of the previous meanings. Thus, the partial meaning five can be considered as an intuitive test of independence, since the onto-semiotic configuration highlighted the notion of association, which subsequent generalization led to the test of independence with the χ^2 statistic.

4.2.1 Partial meaning 5 (PM5): The beginnings of the test of independence through the association coefficient Q

In 1900, Yule was searching for a way to prove an association between two discrete variables, just like the association for the continuous variables was measured. To prove his association coefficient Q , he used a situation/problem where he resumed data from the publication 'Vaccination and Small-Pox Statistics' by Mr. Noel A. Humphrey. He showed a table with the small-pox attack rates in houses that at the moment of the data collection were invaded by small-pox, of persons under and over ten years of age in Sheffield, Warrington, Dewsbury, Leicester, and Gloucester. English cities in which small-pox epidemics have recently occurred. The objective was to show if there is an association between the unvaccinated and the attack in the infected houses of persons under and over such age.

The linguistic elements identified are natural language to formulate the problem, to explain how the test of association is applied and to determine if there is or not association at the time of concluding; tabular representation to provide the small-pox attack rate, of persons under and over ten years of age, in houses affected in towns in which small-pox epidemics have recently occurred; and to indicate the association coefficients. Symbolic representation was also identified in the calculation of the association coefficient.

Regarding the concepts/definitions, some of the most relevant were the variable, qualities or attributes, frequency, sets in probability, correlation coefficient and highlights the one of association (in terms of correlation for continuous variables) and the association coefficient as a measure that approaches the association, towards complete independence on the one hand, and complete association on the other. This coefficient is also understood as a symmetrical function of the attributes, ranging between ± 1 and zero when the attributes are unassociated. The association coefficient $Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$, constitutes a property/proposition. It is also highlighted the property/proposition that initially gives foundation to Yule's elaborations about the association coefficient, augmenting that two qualities or attributes, A and B, are defined as independent if the

chance of finding them together is the product of the chances of finding them separately too, for example, if $\frac{(AB)}{U} = \frac{(A)}{(U)} \cdot \frac{(B)}{(U)}$. Other properties concerning the cross-products were identified when the association coefficient should be $-1, 0, +1$.

The procedures used to find the association measure between the unvaccinated and the attacks in the infected houses of persons under and over the age of ten, through the association coefficient Q , are: (1) forming the 2×2 table of the attributes with which it concerns establishing if there is an association, (2) calculating the association coefficient Q . With the results, it was possible to develop arguments concerning the association coefficient value by age group and about the obtained value of the association coefficients of the different towns.

4.2.2 Partial meaning 6 (PM6): Test of independence through contingency with the χ^2 statistic and the contingency coefficients

In 1904, Pearson proposed a way to measure the contingency that generalized the one of association of Yule (1900). The situation/problem posed by Pearson (1904), was that of the small-pox epidemic of 1890, intending to illustrate with a numerical example his proposal and viability and effectiveness of this over others like Yule's association coefficient.

Regarding the linguistic elements, the ones identified were natural language in the formulation of the problem and the interpretation of results; tabular representation, particularly a 2×2 contingency table in which the observed frequencies were shown. Symbolic representation is also used when showing the results of the calculations of mean square contingency, mean contingency, χ^2 statistic and the association coefficient Q . Concerning these new terms mentioned in the symbolic representation introduced by Pearson, it is necessary to indicate that they are associated with a series of concepts/definitions, properties/propositions, or procedures to prove statistical independence. For example, the concept/definition of contingency generalizes the notion of association of two attributes developed by Yule, and now it is possible to classify individuals not only in two groups but also into as many groups with exclusive attributes as we please and where the order of the sub-groups is of no significance. To measure contingency is necessary to resort to the property/proposition contingency grade, $1 - P$, which indicates that the higher contingency, the higher amount of association between two attributes. This measurement is made through the following procedures: (1) from the values of the contingency table we can find the values of the mean square contingency $\phi^2 = \frac{(ab-cd)^2}{(a+d)(c+b)(a+c)(d+b)}$, $\chi^2 = \frac{(bc-ad)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}$ and mean

contingency $\psi = \frac{2(ab-cd)}{N^2}$; (2) from the values of ϕ^2 , χ^2 and ψ , we can find C_1 , C_2 , Q and the contingency grade (where P is found with the help of the Elderton's tables for goodness-of-fit). Finally, the arguments are made concerning the contingency grade that indicates the deviation of the probabilistic independence that the attributes have. For example, "a case of small-pox and presence or absence of cicatrix is such that the above table could only arise 718 times in 1040 cases if the two events were absolutely independent" (PEARSON, 1904, p. 22). Besides, the contingency coefficients can indicate the degree of association between the attributes.

Other examples of concepts/definitions are the following: (a) table of contingency (a term introduced by Pearson, 1904), formed by s columns and t rows with a total frequency N distributed in sub-groups corresponding to these $s \times t$ compartments; (b) probability, as a measure of how far the observed system is or is not compatible with a basis of independent probability. Other concepts/definitions identified are variable, attributes, observed frequency, expected frequency, mean contingency, association, and correlation. Similarly, other examples of properties/propositions are probabilistic independence; the χ^2 statistic for $s \times t$ and 2×2 contingency tables; mean square contingency; mean contingency; the contingency coefficients; and highlights $n' = (x - 1)(\lambda - 1)/N - \phi^2$, where x are the number of rows and λ the number of columns used for the calculation or search in tables of the probability associated with the χ^2 statistic.

4.2.3 Partial meaning 7 (PM7): Test of independence with degrees of freedom

According to Fisher (1922), for the case of the test of independence with 2×2 contingency tables, we should take $n' = 2$ instead of $n' = 4$, and for the calculation of probability, Elderton's tables can be applied. Fisher resumed a situation/problem with the data used by Greenwood and Yule in 1915, on typhoid, to measure the association between the attacked persons and the vaccine; however, Fisher made the test with the notion of degrees of freedom introduced in 1922.

In Fisher's test of independence, he used linguistic elements such as natural language and tabular representation in the same sense as in PM6. He also used symbolic representation when calculating to obtain the probability through the formula and the value of the statistic. Regarding concepts/definitions, the ones identified were: observed frequency, in the same sense as PM2; variate and probability, following PM6; theoretical frequency, as the number of times that according to a distribution under the assumption of independence a value is expected to occur in a dataset.

Other examples are marginal distribution, which gives one-dimensional information about each classification and says anything about the association between the two variables; and association as a divergence of independence. It was also possible to identify the introduction of degrees of freedom as the number of free parameters minus the number of parameters to estimate. Likewise, this concept was related to one of the critical properties/propositions in this configuration, $k = cr - 1 - (c - 1) - (r - 1) = (c - 1)(r - 1)$, where c indicates the number of columns and r the number of rows. The introduction of this property to the test of independence generated a change in n' , which is now understood as $n' = (c - 1)(r - 1) + 1$.

The procedures identified are: (1) calculating the χ^2 statistic through the formula proposed by Fisher (1922), for a 2×2 contingency table, $\chi^2 = \frac{(bc-ad)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}$, in case of working with a table $c \times r$ it is necessary to obtain the expected frequencies under the assumption of probabilistic independence from observed frequencies through the observed marginal totals; (2) obtaining n' ; and (3) looking for the probability value in the table of the χ^2 distribution or in the table of critical values of Fisher.

The arguments used by Fisher consisted in the probability associated with the χ^2 statistic and the comparison with the predetermined limit as significant deviation $P = 0.05$. In his example, since $\chi^2 = 56.234$, Fisher indicated that the observations were opposed to the hypothesis of independence.

4.2.4 Partial meaning 8 (PM8): Test of independence with Yates' continuity correction factor

In this partial meaning, the problem of PM5 is resumed, which refers to Yates' correction for continuity that the χ^2 statistic requires when working with small numbers. This continuity correction emerges in the context of the test of independence in 1934. Yates wanted to test how the correction that he proposed affects the work with frequencies under five. For this, he resumed the situation/problem and the data of Dr. Milo Hellman that referred to malocclusion, which are alterations in the position of infant's teeth, and how the baby was fed. Hellman concludes that bottle-feeding is one of the factors causing malocclusion.

Regarding the linguistic elements that Yates used, it was possible to identify natural language and tabular representation in the same sense as in PM6, also the probability of the number of breastfed infants is shown in the tabular representation. Symbolic representation is also used in the calculation to obtain the statistical value, the statistic with correction for continuity, and

the probability distribution of the number of breastfed children. Concepts/definitions from other partial meanings are resumed, for example, variable from PM6; observed frequency, expected frequency, frequency distribution, marginal distribution, and probability, from PM7; while the correction for continuity factor is understood in the same sense as in PM4. Concerning the concept/definition of independence between two variables, indicates that there is no relation (association) between the variables under study; thus, such variables have a distribution under probabilistic independence.

Properties/propositions from PM7 are also resumed, particularly, the degrees of freedom and n' . The statistic with correction for continuity factor is introduced under the expression $\chi_c^2 = \sum \frac{\left(\left(a - \frac{1}{2} \right) \left(d - \frac{1}{2} \right) - \left(b + \frac{1}{2} \right) \left(c + \frac{1}{2} \right) \right)^2}{n \times n' (N - n) (N - n')}$, for 2×2 contingency tables. This last property corresponds to the first action of a series that conform the procedures to find the probability of occurrence of the event when the χ_c^2 statistic is used to test independence between two variables of the same population. The probability of occurrence of the event is used to propose the arguments. Yates also indicated that it was possible to argue about the value of the statistic and the critical value of the distribution.

4.3 Problem area 3: test of homogeneity

In this section, we will describe the characteristics of the four partial meanings identified in the historical evolution of the test of homogeneity with the χ^2 statistic.

4.3.1 Partial meaning 9 (PM9): The beginnings of the test of homogeneity with the χ^2 statistic

In 1911, Pearson wanted to find out the probability that two samples known a priori may be from the same population, however not having a priori knowledge of the population. To show how the test of homogeneity with the χ^2 statistic work, Pearson resumed a situation/problem with data from Dr. Macdonald's studies, with which he pretended to determine if there was a selective disease. Mainly, he wondered whether hair color exercises a differential selection regarding the incidence of scarlet fever and measles. He applied the test of hair color with two samples; the first sample corresponds to persons with scarlet fever and the second to persons with measles.

The linguistic elements that Pearson used to make the test of homogeneity are natural language and tabular representation in the same sense as in PM6, symbolic representation is also

used in the table of contingency and in the calculation to obtain the value of the statistic and when referring to probability. Concerning the concepts/definitions introduced on this partial meaning, we have independent samples, which are samples selected randomly so that its data do not depend on other observations; probability, which in this context is understood as a measure of occurrence that both are random samples of the same population. Likewise, concepts/definitions from PM6 are resumed, for example, variable, attribute, or class, and observed frequency.

One of the properties/propositions from which Pearson substantiated the use of the χ^2 statistic for the test of homogeneity refers to whether the samples are independent then $\sigma_p^2 = n^2 \left(\frac{\Sigma_p^2}{N^2} + \frac{\Sigma_p'^2}{N'^2} \right)$, $\sigma_p \sigma_q r_{pq} = n^2 \left(\frac{\Sigma_p \Sigma_q R_{pq}}{N^2} + \frac{\Sigma_p' \Sigma_q' R_{pq}'}{N'^2} \right)$, being $\Sigma_p, \Sigma_q, \Sigma_p', \Sigma_q'$ the standard deviations of the frequencies of the p th, q th frequencies of the two samples and R_{pq}, R_{pq}' the correlations of the same frequencies. Additionally, if the frequencies belong to random samples of the same population, we have $\frac{\bar{f}_p}{N} = \frac{f_p'}{N'} = \frac{\mu_p}{M}$, $\frac{\bar{f}_q}{N} = \frac{f_q'}{N'} = \frac{\mu_q}{M}$. As it was mentioned, the previous properties/propositions substantiated the following, $\chi^2 = S_1^g \left\{ \frac{NN' \left(\frac{f_p}{N} - \frac{f_p'}{N'} \right)^2}{f_p + f_p'} \right\}$. Calculating the

value of this property also constitutes the first action of the procedures to find the probability that the two samples derive from the same population, to find this probability he was assisted by Elderton's tables of goodness-of-fit (χ^2 and n'), and is precisely this probability the one used to make the arguments to justify and return to the context of the problem addressed. Pearson's arguments about the test of homogeneity indicated that: "The odds are more than 33,000 to 1 against the occurrence of two such divergent samples of hair color if they were random samples from the same population" (PEARSON, 1911, p. 252).

4.3.2 Partial meaning 10 (PM10): Snedecor's test of homogeneity

Snedecor and Irwin (1933), presented a proposal for the homogeneity test that would apply to the results of experiments, where the frequencies of the observations that arise from different subsets are unequal. The type of situations/problems with which they worked referred to laboratory experiments, where the frequencies of the observations derived from different unequal subsets, for example, the mortality in epidemics induced to laboratory animals and controlled infestations in croplands.

The linguistic elements characterized were natural language, in the same sense as in PM6, and tabular representation to provide the number of apples of each sub-sample, number of injured apples, and percentage of the sub-sample to which the injured apples correspond for both fumigation methods. Symbolic representation is also used to present the values of the statistic, the probability in the table of contingency, and in the conclusions. The concepts/definitions are variable resumed from PM6; observed frequency from PM7; and population and sample from PM9. However, in this case, the sample is composed of the combination of diverse sub-samples taken from different places. Two relevant concepts/definitions in this configuration are probability and homogeneity. The probability is understood as a measure of occurrence that the series of sub-samples have been extracted from a homogeneous population, in the sense that the probability of the event is uniform throughout the experimental material. In contrast, homogeneity refers to a measure to see if it is possible that the samples come from a homogeneous population.

Regarding the essential properties/propositions for this test, we have the degrees of freedom $n = (c - 1)(r - 1)$, the probability of the attribute in each sub-sample $p = \frac{100s}{n} \%$, the average probability for the whole sample $\bar{p} = \frac{100 \sum s}{\sum n} \%$, the average probability of the complement of the attribute $\bar{q} = 100 - \bar{p} \%$, and the statistic $\chi^2 = \frac{100 (\sum sp - \bar{p} \sum s)}{\bar{p} \bar{q}}$, in which $\sum sp$ is the sum of the products of the pair of the frequency of the attribute in each sub-sample.

The procedures identified are: (1) using the property/proposition $\chi^2 = \frac{100 (\sum sp - \bar{p} \sum s)}{\bar{p} \bar{q}}$; (2) establishing the number of degrees of freedom; and (3) looking for the value of the probability in the table of the χ^2 distribution for the value of χ^2 and n . With this probability obtained from the tables, it is possible to make the arguments to justify the results of the test of homogeneity of the population.

4.3.3 Partial meaning 11 (PM11): Test of homogeneity with the degrees of freedom and Fisher's χ^2 distribution table

According to Fisher (1934), the tests of homogeneity and independence with the χ^2 statistic are mathematically identical; hence it is also resumed from PM7 the fact that with 2×2 contingency tables we should take $n' = 2$ instead of $n' = 4$. Fisher (1934), showed how the test of homogeneity works with the χ^2 statistic, for which he resumes a problem with samples of different sizes about data of crustaceans that he had previously analyzed at the request of Huxley.

The test aimed to study the homogeneity of different families for the black and red eyes proportion.

Regarding the linguistic elements, the ones used were natural language, in the same sense as in PM6, tabular representation to show the eye color frequency per family. Symbolic representation was also used in the calculation to obtain the value of the statistic and the probability. The concepts/definitions are resumed from different partial meanings, for example, variable from PM6; observed and expected frequency, frequency distribution and marginal distribution from PM7; and sample, probability, and homogeneity from PM9. Concerning the properties/propositions, the ones introduces are: Expected frequency as $e_{ij} = \frac{n_i \times n_j}{n}$, degrees of freedom $n = (c - 1)(r - 1)$, and when $n > 30$ the expression used is $(\sqrt{2\chi^2} - \sqrt{2n - 1})$, if the value of the expression is over two, the value of χ^2 is not following the expectations. Regarding the χ^2 statistic, it is understood in the same sense as in PM7.

The procedures identified are: (1) calculating the statistic through $\chi^2 = \sum \frac{((m+x)-m)^2}{m}$, to which is necessary to obtain the expected frequencies under the assumption of independence from observed frequencies through the marginal totals observed $e_{ij} = \frac{n_i \times n_j}{n}$; (2) obtaining the number of degrees of freedom n ; and (3) using the expression $\sqrt{2\chi^2} - \sqrt{2n - 1}$, since $n > 30$, and it is not found within the tabulated values. The arguments focused on the value of the expression $\sqrt{2\chi^2} - \sqrt{2n - 1}$, which is below two and indicated that "the series is therefore not significantly heterogeneous; effectively all the families agree and confirm each other in indicating the black-red ratio observed in the total" (FISHER, 1934, p. 95).

4.3.4 Partial meaning 12 (PM12): Test of homogeneity with Yates' continuity correction factor

To show how the test of homogeneity works with the correction for continuity factor to the χ^2 statistic, we resume the problem presented by Pearson (1911) about the hair color of two samples. The first sample corresponds to persons with scarlet fever and the second sample to persons with measles.

Among the linguistic elements in this configuration, we have natural language, in the same sense as in PM6, and tabular representation to show the frequencies of hair color of two samples (just as in PM11). Likewise, symbolic representation is also used as in PM11. Concerning the concepts/definition and the properties/proposition, they are resumed from different partial meanings, for example, in concepts/definitions the correction for continuity factor is resumed from

PM4; variable from PM6; observed and expected frequency, frequency and marginal distribution, from PM7; and sample, probability, and homogeneity from PM9. Concerning properties/proposition, the degrees of freedom and expected frequency are resumed from PM11. Moreover, the χ^2_c statistic is understood as in PM8.

The procedures identified are: (1) calculating the expected frequencies and the value of the χ^2_c statistic; (2) establishing the degrees of freedom; and (3) looking for the value of the probability in χ^2 distribution tables. While the arguments revolved around the probability of occurrence that both samples correspond to the same population. For example, for $\chi^2 = 26$ and $n = 4$, the χ^2 distribution table shows $P = 0.000,086039$; hence, the probabilities are around 11,628 to 1 against the emergence of two samples of hair color that diverge, if they were random samples from the same population.

4.4 Problem area 4: Chi-square distribution

Karl Pearson was the first obtaining the χ^2 distribution as the asymptotic distribution of the χ^2 statistic (HEYDE; SENETA, 1977). Pearson (1900), made that connection when he was working the problem of goodness-of-fit for a frequency curve. Earlier, the evolution of the χ^2 statistic to solve diverse types of problems, and the application of the χ^2 distribution to solve such problems has been described and characterized. Within the variations of the distribution application, it has been addressed what is recognize as degrees of freedom and its importance to determine the probability and the critical regions in the χ^2 distribution. Furthermore, it is also recognized the importance that the χ^2 distribution tables had and that they facilitated the use of the contrast tests of the χ^2 statistic and it was Pearson (1900) who provided the first version of them, in which it was possible to find the probability with (χ, n') , establishing the formulas to calculate the probability in the distribution to any n' . Subsequently, Elderton (Pearson, 1914) also published an extension of such tables for the probability (χ^2, n') . Fisher (1934), also published tables for the distribution with (n, P) , where χ^2 is obtained to establish the regions of acceptance and rejection, and established an expression for the calculation of the critical value of χ^2 in the distribution for the cases of $n > 30$.

Pearson (1900), renamed his Type III curve as ' χ^2 distribution'; at that moment, he does not refer to Helmert, who had been working with such distribution, and it is not until 1931 that Pearson recognizes the precedent of the distribution. In the following paragraphs, some of the precedents of the distribution are presented.

According to Lancaster (1966), Pearson's χ^2 distribution can be considered as the culmination of Laplace's work on least squares. Laplace worked with the gamma distribution in the context of the theory of errors, being the χ^2 distribution a case of the gamma distribution. Laplace provided with his work the necessary techniques for the subsequent Bienaymé's work of 1838, where he developed the distribution in a linear form in the success and failures of the binomial as preliminary to an extension of the multinomial theory; and in 1952, he obtained the χ^2 distribution as an asymptotic result without the assumption of normality (LANCASTER, 1966). Lancaster also indicated that in 1844 Ellis worked in a problem suggested by Isaac Newton on reigns length, from which he provided a method to determine the distribution of the sum of n independently distributed random variables by the use of Laplace's transforms.

According to Lancaster (1966) and Hald (2007), in 1875 Helmer provided a frequency distribution of the sum of squares $\sum \varepsilon_i^2 = n\sigma^2$, where ε_i are mutually independent normal variables, with an expected value equal to zero, where the expression is $f\left(\frac{v}{2}\right) = \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{v}{2}\right)^{\frac{n}{2}-1} \exp\left(-\frac{v}{2}\right)$, then v has a χ^2 distribution with n degrees of freedom. In 1876, Helmer replaced n for $n - 1$. Helmer (1876), considered joint distribution of the z_i and applied maximum likelihood and obtained $\hat{\sigma}^2 \sum z_i^2/n$, then obtained the joint distribution of the $n - 1$ differences $u_j = x_j - \bar{x}$. He also made a transformation to $\sum (x - \bar{x})^2/\sigma^2$ for the x 's and the v 's, which is known as Helmer transformation, which showed that $\sum_1^n (x_i - \bar{x})^2/\sigma^2 = \sum_1^{n-1} v_i^2/\sigma^2$ is distributed as χ_{n-1}^2 , where $x_i = \mu + z_i$.

5 Incidence of partial meanings identified for the statistical education research

In the present section, we describe the connections among partial meanings identified in this article, for the χ^2 statistic, and the developments of the statistical education research concerning IIR and FIR, the use of statistical software, and the textbooks.

In this sense, diverse authors (e.g., Woodard et al. 2020; Fellers and Kuiper 2020; DePaolo et al. 2016; Seier 2014; Gibbs and Goossens 2013; Leigh and Dowling 2010), have reported activities or problems proposals with the finality to make more accessible the study of the χ^2 tests, goodness-of-fit, independence, and homogeneity, which correspond to the problem areas presented in the previous section. However, commonly, in these investigations, the last partial meanings of each problem area (i.e., PM3, PM7 y PM11) were used. That is to say; problems are proposed to promote the FIR by students, due to the approach used has a high degree of

formalization (although these authors do not explicitly refer to FIR). We consider that the first partial meanings of each problem area identified here, as in the case of PM1, could be used intuitively to promote the logic of the goodness-of-fit-test with the χ^2 statistics, that is to say, promote the IIR for the case of this test.

For example, Leihg and Dowling (2010) present data that they collected about the water taste from different brands and how these data could be used to perform inferential analyzes with the χ^2 goodness-of-fit and independence test. For the case of the goodness-of-fit-test, primary mathematical objects corresponding to PM3 can be identified, such as the followings procedures: establish the degrees of freedom (also associated with the proposition/property), calculate the χ^2 statistic and the p-value. Moreover, linguistic elements such as the use of natural language and the usual tabular representation in this type of test are identified.

In the same sense, Gibbs and Goossens (2013) propose to analyze a dataset about the efficacy of HPV vaccines with the χ^2 homogeneity test. An analysis of the activity statement and solution, in terms of the meanings proposed here, allow revealing the primary mathematical objects that correspond to PM11. Another example is that of DePaolo et al. (2016), who propose an activity about a sample of 200 transactions, which is solved with the χ^2 independence test, where primary mathematical objects used (and described) corresponding to PM7.

In these last two works, software was used for statistical data analysis; in the case of the first study, the authors point out that they do not use the 'continuity correction factor,' although some data have a frequency of less than five. This is very usual when we are working with software because some (e.g., Minitab) do not include the option to use the continuity correction, which constitutes a concept/definition and property/proposition corresponding to PM4, PM8, and PM12. When we are working with frequencies less than five, the χ^2 tests require this correction, but the statistical software often gives a warning as an additional note to the results. Therefore, another use that can be made of the identified meanings in this article refers to promote what is found in the background understanding in the processes or procedures carried out by some statistical software, in this case for χ^2 tests, this would help a critical review of the statistical results produced by the software, with the aim of interpreting them correctly.

Furthermore, university textbooks (e.g., Devore 2008), present activities for the χ^2 tests, goodness-of-fit, independence, and homogeneity; however, the tests are introduced in an 'abrupt' way, completely formalized and without an introduction or progression in the understanding of the test. For example, the first activity on the goodness-of-fit-test in Devore (2008), is about the

phenotypes from a dihybrid cross of tomatoes and requiring probing if they are distributed according to Mendel's laws. An analysis considering the epistemological characteristics of the identified partial meanings would allow us to observe that the practices developed to perform the goodness-of-fit-test correspond to PM3. While the first activity with the presentation of the homogeneity test deals with the homogeneity in production lines concerning five categories that make defective products. In a brief analysis, it was possible to identify concepts/definitions such as frequency distribution and marginal distribution, properties/propositions like the expected frequency and degrees of freedom, procedures like calculating the statistic, for which it is necessary to calculate the expected frequency. The previous primary mathematical objects are used in the sense of PM11. Finally, the analysis of the activity for the independence test revealed that primary mathematical objects used for the development of this test correspond to PM7.

The analyzes presented could be carried out in greater depth, just as we did with the situations identified in the historical evolution. However, for space reasons and not being the scope of this work, they are presented succinctly. The objective of evidencing them is to evince the uses that can be given to partial meanings and to primary mathematical objects associated with them, as well as, to analyze, and to propose activities for promoting the meanings progressively (from informal to formal) of the χ^2 statistic in teachers and students.

In this sense, this work purports to be the first step towards a proposal of progressive levels of inferential reasoning (that allow a continuous transition from IIR to FIR) for the χ^2 statistic, in other words, a progression from intuition towards formalization. Indeed, a proposal like this would lead to revisiting discussions of statistical education research that lead to associating the IIR with, almost exclusively, descriptive statistics concepts (see, for example, Zieffler et al. 2008; Makar and Rubin 2009). Thus, a line of research that is generated from our study is related to 'a new perspective' of the IIR, perhaps understanding this type of reasoning as processes that involve different epistemological characteristics of partial meanings (as well as, intuitive, as pre-formal). In other words, it may be necessary to rethink what we understand by IIR and FIR, in terms of mathematical and statistical objects and processes associated with the meanings of the notions that we are working in our research area.

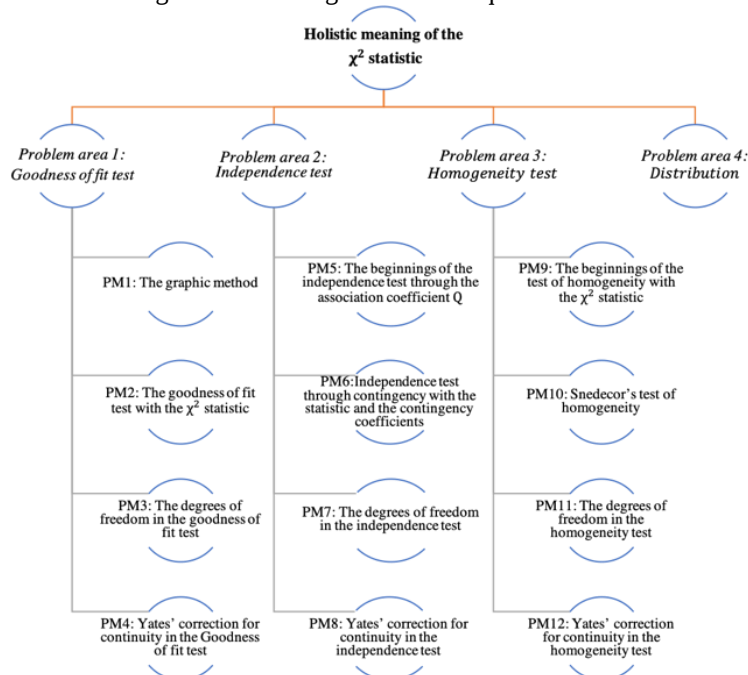
6 Final reflections

The present article was intended to characterize the meanings attributed to the Chi-square statistic throughout history. With this aim, and with the support of notions from the onto-semiotic

approach (OSA), it was possible to identify four crucial problems that, when addressed, give rise to four broad "meanings of reference" for this statistic: 1) Goodness-of-fit-test; 2) Test of independence; 3) Test of homogeneity; 4) Chi-square distribution. Furthermore, each of the broad meanings of reference is composed of partial meanings, which show the progressive evolution (from informal to formal) of the χ^2 statistic (figure 2).

An important aspect worth clarifying, although implicitly done throughout this article, it is the emphasis that we give to the χ^2 statistic and not to the historical development of the χ^2 distribution, this is due to the fact that with the emergence and development of the χ^2 statistic, establishing the χ^2 distribution as the asymptotic distribution of the χ^2 statistic and the application of the test of independence set off refinements in the χ^2 distribution, particularly as regards to the degrees of freedom. It is recognized that this distribution has a crucial role because it is the distribution that the χ^2 statistic follows and for the refinements that the 1, 2 and 3 problems had with the evolution of the χ^2 distribution.

Figure 2: Meanings of the Chi-square statistic



Source: prepared by the authors

We consider that the present characterization of the meanings of the χ^2 statistic enables us to access to the mathematical richness and visualize the variety of paths to the teaching and learning of this notion, which clearly constitutes an advance for the scientific community, and training of teachers, interested in the teaching of statistical notions like the χ^2 statistic. While the meanings characterized in this article permit to identify mathematical-statistical elements for a continuous and progressive transition from IIR to FIR, which could serve, as a future line of

research, to build progressive levels from informal to formal, of inferential reasoning on the χ^2 statistic. An example of the use of mathematical-statistical elements for the construction of progressive levels of inferential reasoning on the t-Student statistic can be seen in Lugo-Armenta and Pino-Fan (2021). The elements of the configuration associated to each one of the meanings respectively, constitute epistemic guidelines that will enable to build tasks or didactic sequences that promote this type of reasoning with diverse meanings of the χ^2 statistic.

7 Acknowledgment

This article has been developed within the framework of the research project Fondecyt 1200005, funded by Agencia Nacional de Investigación y Desarrollo (ANID) of Chile.

8 References

- BARNARD, George Alfred. 1992. Introduction to Pearson (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. In: KOTZ, S.; JOHNSON, N.L. (Eds.), **Breakthroughs in Statistics**, v. 2. New York: Springer, 1992. p. 1-10.
- BATANERO, Carmen. Del análisis de datos a la inferencia: Reflexiones sobre la formación del razonamiento estadístico. **Cuadernos de Investigación y Formación en Educación Matemática**, v. 8, n. 11, p. 227-291, dic. 2013.
- BAKKER, Arthur.; GRAVEMEIJER, Koeno. Learning to reason about distribution. In: BEN-ZVI, Dani; GARFIELD, Joan. (Eds.), **The challenge of developing statistical literacy, reasoning, and thinking**. Dordrecht: Kluwer Academic Publishers, 2004. p. 147-168.
- COHEN, Louis.; MANION, Lawrence.; MORRISON, Keith. **Research methods in education**. London and New York: Routledge, 2011.
- DEPAOLO, Concetta A.; ROBINSON, David F.; JACOBS, Aimee. Café Data 2.0: New Data From a New and Improved Café. **Journal of Statistics Education**, v. 24, n. 2, p. 85-103, jun. 2016.
- DEVORE, Jay L. **Probability & Statistics for Engineering and the Sciences**. 7. ed. Mexico: Cengage Learning, 2008.
- DOERR, Helen M.; DELMAS, Robert; MAKAR, Katie. A modeling approach to the development of students' informal inferential reasoning. **Statistics Education Research Journal**, Auckland, v. 16, n. 2, p. 86-115, nov. 2017.
- FELLERS, Pamela S.; KUIPER, Shonda. Introducing Undergraduates to Concepts of Survey Data Analysis. **Journal of Statistics Education**, v. 28, n. 1, p. 18-24, feb. 2020.
- FISHER, Ronald Aylmer. On the interpretation of χ^2 from contingency tables, and the calculation of P. **Journal of the Royal Statistical Society**. v. 85, n. 2, p. 87-94, jan. 1922.

FISHER, Ronald Aylmer. **Statistical methods for research workers**. 5. ed. Edinburgh: Oliver and Boyd, 1934.

GALTON, Francis. IV. Statistics by intercomparison, with remarks on the law of frequency of error. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 49, n. 322, p. 33-46, 1875.

GALTON, Francis. The Application of a Graphic Method to Fallible Measures. **Journal of the Statistical Society of London**. p. 262-271, jun. 1885.

GIBBS, Alison L.; GOOSSENS, Emery T. The Evidence for Efficacy of HPV Vaccines: Investigations in Categorical Data Analysis. **Journal of Statistics Education**. v. 21, n. 3, 2013.

GODINO, Juan Díaz; BATANERO, Carmen. Significado institucional y personal de los objetos matemáticos. **Recherches en Didactique des Mathématiques**, Grenoble, v. 14, n. 3, p. 325-355. 1994.

GODINO, Juan Díaz; BATANERO, Carmen; FONT, Vicenç. The onto-semiotic approach to research in mathematics education. **ZDM**, Berlin, v. 39, n. 1-2, p. 127-135, mar. 2007.

GODINO, Juan Díaz; BATANERO, Carmen; FONT, Vicenç. The onto-semiotic approach: implications for the prescriptive character of didactics. **For the Learning of Mathematics**, Vancouver, v. 39, n. 1, p. 37-42. 2019.

GODINO, Juan Díaz; FONT, Vicenç; WILHELMI, Miguel R.; LURDUY, Orlando. Why is the learning of elementary arithmetic concepts difficult? Semiotic tools for understanding the nature of mathematical objects. **Educational Studies in Mathematics**. v.77, p. 247-265, jul. 2011.

GREENWOOD, Major; YULE, George Undy. The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general. **Proceedings of the Royal Society of Medicine**, v. 8, p. 113-194, jun. 1915.

HALD, Anders. **A history of parametric statistical inference from Bernoulli to Fisher, 1713-1935**. Springer Science & Business Media, 2007.

HELMERT, Friedrich Robert. Über die Wahrscheinlichkeit der Potenzsummen der Beobachtungsfehler und über einige damit im Zusammenhange stehende Fragen. **Z. Math. und Physik**, v. 21, p. 192-218, 1876.

HEYDE, Chris; SENETA, Eugene. **I. J. Bienaymé: Statistical Theory Anticipated**. New York: Springer, 1977.

HOEKSTRA, Rink. Risk as an Explanatory Factor for Researchers' Inferential Interpretations. **The Mathematics Enthusiast**, v. 12, n. 1, p. 103-112, 2015.

JACOB, Bridgette. L.; DOERR, Helen. M. Statistical Reasoning with the sampling distribution. In: MAKAR, Katie; DE SOUSA, B.; GOULD, R. (Eds.), **Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics**. Voorburg, The Netherlands: International Statistical Institute, 2014. p. 1-6.

LANCASTER, H.O. Forerunners of the Pearson χ^2 . **Australian Journal of Statistics**, n. 8, n. 3, p. 117-26, nov. 1966.

LEIGH, M.; DOWLING Alix .D. Water Taste Test Data. **Journal of Statistics Education**, v. 18, n. 1, 2010.

LUGO-ARMENTA, Jesús Guadalupe; PINO-FAN, Luis Roberto. Niveles de Razonamiento Inferencial para el Estadístico t-Student. **Bolema: Boletim de Educação Matemática**, v. 35, n-71, en prensa, 2021.

MAKAR, Katie; RUBIN, Andee. A framework for thinking about informal statistical inference. **Statistics Education Research Journal**, Auckland, v. 8, n. 1, p. 82–105, may. 2009.

MAKAR, Katie.; RUBIN, Andee. Learning about statistical inference. In: BEN-ZVI, D.; MAKAR, K.; GARFIELD, J. (Eds.), **International handbook of research in statistics education**. Switzerland: Springer International. 2018. p. 261-294.

MAGNELLO, M. Eileen. Karl Pearson, paper on the chi square goodness of fit test (1900). In **Landmark Writings in Western Mathematics 1640-1940**. Elsevier Science, p. 724-731, 2005.

PEARSON, Karl. 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**. v. 50, n. 5, p. 157-175, 1900.

PEARSON, Karl. **On the theory of contingency and its relation to association and normal correlation**. Dulau and Company, 1904.

PEARSON, Karl. On the probability that two independent distributions of frequency are really samples from the same population. **Biometrika**, v. 50, n. ½, p. 250-254, jul. 1911.

PEARSON, Karl. **Tables for Statisticians and Biomnetricians**. Cambridge: University Press, 1914.

PFANNKUCH, Maxine.; ARNOLD, Pip; WILD, Chris. J. What I see is not quite the way it really is: Students' emergent reasoning about sampling variability. **Educational Studies in Mathematics**, New York, v. 88, n. 3, p. 343-360, mar. 2015.

PINO-FAN, Luis Roberto; FONT, Vicenç; GORDILLO, Wilson; LARIOS, Victor; BREDÁ, Adriana. Analysis of the meanings of the antiderivative used by students of the first engineering courses. **International Journal of Science and Mathematics Education**, v. 16, n. 6, p. 1091-1113, 2018. DOI: <https://doi.org/10.1007/s10763-017-9826-2>.

PINO-FAN, Luis Roberto; GODINO, Juan Díaz; FONT, Vicenç. Epistemic Facet of the Didactic-Mathematics Knowledge About The Derivative. **Educação Matemática Pesquisa**, v. 13, n. 1, p. 141-178, 2011.

REABURN, Robyn. Introductory statistics course tertiary students' understanding of p-values. **Statistics Education Research Journal**, v. 13, n. 1, p. 53-65, may 2014.

RIEMER, Wolfgang; SEEBACH, Günter. Rolling pencils - inferential statistics in the pencil case. In **Understanding more mathematics with GeoGebra**. Heidelberg: Springer Spektrum, p. 91-105, 2014.

ROCHOWICZ, John A. Bootstrapping analysis, inferential statistics and EXCEL. **Spreadsheets in Education (eJSiE)**, v. 4, n. 3, p. 1-23, 2010.

ROSSMAN, Allan J. Reasoning about Informal Statistical Inference: One Statistician's View. **Statistics Education Research Journal**, Auckland, v. 7, n. 2, p. 5-19, nov. 2008.

SALDANHA, Luis A.; THOMPSON, Patrick W. Conceptions of sample and their relationship to statistical inference. **Educational Studies in Mathematics**, v. 51, n. 3, p. 257-270, nov. 2002.

SEIER, Edith. An Early Start on Inference. In I. GAL (ed.), **9th International Conference on Teaching Statistics**. Conference held in Flagstaff, Arizona, United States of America, 2014.

SNEDECOR, G.; IRWIN, M.R. On the chi-square test for homogeneity. **Iowa State Coll. J. Sci**, v. 8, n. 1, p. 75-81, 1933.

TARLOW, Kevin R. Teaching principles of inference with ANOVA. **Teaching Statistics**, v. 38, n. 1, p. 16-21, 2016.

WIJAYATUNGA, Priyantha. Geometric view on Pearson's correlation coefficient and a generalization of it to non-linear dependencies, **Ratio Mathematica**, v. 30, p. 3-21, 2016.

WOODARD, Victoria; LEE, Hollylynne; WOODARD, Roger. Writing Assignments to Assess Statistical Thinking, **Journal of Statistics Education**, v. 28, n. 1, p. 32-44, 2020.

YATES, Frank. Contingency tables involving small numbers and the χ^2 test. **Supplement to the Journal of the Royal Statistical Society**, v. 1, n. 2, p. 217-35, 1934.

YULE, George Undy. On the association of attributes in statistics: with illustrations from the material of the childhood society, &c. **Philosophical Transactions of the Royal Society of London**. Series A, v. 194, p. 257-319, 1900.

ZIEFFLER, Andrew; GARFIELD, Joan; DELMAS, Robert; READING, Chris. A framework to support research on informal inferential reasoning. **Statistics Education Research Journal**, Auckland, v. 7, n. 2, p. 40-58, nov. (2008).