Universal Journal of Educational Research 4(12): 2687-2700, 2016 http://www.hrpub.org
DOI: 10.13189/ujer.2016.041201

Interpretation of Confidence Interval Facing the Conflict

Luisa Andrade*, **Felipe Fernández**

Department of Mathematics, National Pedagogic University, Colombia

Abstract As literature has reported, it is usual that university students in statistics courses, and even statistics teachers, interpret the confidence level associated with a confidence interval as the probability that the parameter value will be between the lower and upper interval limits. To confront this misconception, class activities have been designed with the aim of realizing that this application of confidence level violates the basic laws of probability, when considering two non-overlapping confidence intervals, that could plausibly correspond to two random samples from the same population, where the probability of events within this interpretation contradicts the probability rule for disjoint events and the rule of monotonicity ($P[E] \leq P[F]$ if $E \subset F$). Afterwards, we use simulation to help students shift to a frequentist interpretation of confidence intervals. Although the expected questioning was generated in students, it does not look like it is enough to establish a solid re-conceptualization of confidence level. We believe that this is due in part, to the slight language used in the teaching process and to incipient conceptions about the probability notion.

Keywords Confidence Interval, Confidence Level, Frequentist Interpretation, Probability Conceptions, Probability Rules, Class Activities

1. Introduction

The research work in a recent study[1] (see e.g. Andrade, Fernández & Álvarez [1]) has focused on the meaning of the confidence level associated with a confidence interval. It is possible to find wide research about identification of misconceptions concerning those notions, and experts point out that both, students and teachers, reveal such misunderstandings, however the references to research design that propose the planning or implementation of instruction aimed at trying to overcome them, are rare. In this paper we present class activities that have been designed, implemented and monitored as part of the study mentioned, with the intention of confronting and provoking a conceptual change in preservice mathematics teachers[2], of the common interpretation for confidence level associated with a confidence interval: the probability that the parameter value will be between the lower and upper interval limits. Also, we introduce a rationale for the activities that include the learning assumptions that underlie them. Then, we discuss results that identify students' conceptions about the notion of confidence level and show the transitions that occur during the developed activities, along with some considerations that may shed light on learning and teaching confidence intervals.

2. Research Methodology

Initially, the study was framed within a qualitative instructional design paradigm that pursuits to expand the roles of researchers and teachers: the first become direct observers of student work, and the second collaborate on creating the research strategies proposed for the class [2, 3]. Under this approach, we decided to work with teaching experiments, which are not intended to evaluate the effectiveness of a previously planned instructional design, but, as noted by Cobb [4], to dynamically construct and improve a design, as long as conjectures on student learning, are tested and modified.

Later on, we embrace working with the methodology of case studies along with teaching experiments, in order to limit the number of students so it would make possible a detailed and accurate monitoring of their work and progress, in coincidence with Neiman & Quaranta [5], whom recognized that case studies focus on "limited number of events and situations to be addressed with the required depth for its holistic and contextual understanding", and specifically with Ponte [6], when he indicates that case studies are used to investigate among other things, students learning issues and aims "to fully understand the hows and whys" of this situation. In addition, we agree with Bogdan &

[1] A research project developed in Universidad Pedagógica Nacional, Bogotá, Colombia, during 2013 and 2014.

[2] College students (preservice teachers) from Universidad Pedagógica Nacional, Bogotá, Colombia.

Biklen (1982, cited in Colas & Buendia [7]) who state that although case studies are a common strategy of qualitative research characterized by a detailed and intensive examination of a situation, a subject or an event, its projection, as in any qualitative research, it is not to get universal abstractions but concrete and specific universals through examination and comparison of case studies. Thus, we selected three students as case studies, renamed Carlos (**C**), Alejandro (**A**) and Bibiana (**B**), who sometimes work in pairs with colleagues. Though the main core of the project results refers to these case studies, there are some general results coming from the whole students group that help to confirm our elucidations.

3. Confidence Intervals and Their Confidence Level

A confidence interval is a type of inference for estimating the value of a population parameter based on the sampling distributions of statistics; it can also be seen as a way of describing the reliability of the estimation based on a point estimator. In other words, when data are not available for the entire population, the confidence interval calculated from sample data is used as an estimation, which provides more information than the sample mean, because it procures a range of values at each side of this mean. In this sense, Davies [8] expresses that confidence intervals add information about the extent of statistic's effects. Under equal circumstances, a larger sample will give a better estimation of the parameter; but also a wider interval is determined by a higher confidence level and therefore it has more chance of success, while a smaller interval with the same confidence level offers a more accurate estimation, but with less chance of success.

The probability of success in the estimation by a confidence interval is usually represented as '$1 - \alpha$' or '$100\times(1 - \alpha)\%$' and is called confidence level where α is the random error or the significance level, that is, the measure of the chance of failure in the estimation by this interval; based on this consideration, if P is the probability distribution function of estimator θ, $P(l \leq \theta \leq u) = 1 - \alpha$, where l and u are the lower and upper interval limits respectively. For a confidence interval is asserted that with a given probability, it is one of the intervals that contains the value of an unknown population parameter. This means that the confidence level indicates the probability that one of the intervals produced in the process to generate confidence intervals, contains the parameter. For example, a 90% confidence interval is one of 90 of 100 possible intervals obtained under the same sampling conditions, which capture the population mean, or, if estimation by confidence intervals of 90% is used several times, it is expected that the population mean is in 90% of the intervals computed. Then, as Behar [9] says, probability refers to the likelihood of the method for computing intervals and not to the parameter; if sampling is repeated a sufficient number of times, the percentage of intervals generated, that captures the population parameter, is given by the confidence level as well. This author emphasizes the meaning of 'confidence' as "the potential repetition of the estimations, so that the specific interval obtained as a result of a random sample does not have the 'confidence'; instead the confidence is associated with the random process that generates the interval". In the same direction, Montgomery & Runger [10] suggest that the process of estimation by confidence intervals ensures that the method used to construct the interval, produces trustworthy statements about the interval containing the parameter, $100\times(1 - \alpha)\%$ of the times, or in words of Freund & Walpole [11] "if we had to bet, 95-5 would be fair chance" that the parameter is between the interval limits. Thus, it is clear that the interpretation of confidence level, as researchers who have been studying it advert us, must incorporate the probability frequentist perspective for the confidence interval generated.

Understanding confidence interval requires knowledge of the mathematical objects related with it, like population, sample, statistic, parameter, standard error, sampling distribution, critical value, and to be familiarized with the theoretical models of sampling distribution, in order to be able to determine the particular model of the statistic distribution associated with the parameter θ that will be estimated. It is common that this distribution approaches normal or t-student distribution models. In theory, a confidence interval of $100\times(1 - \alpha)\%$, built to estimate a population parameter θ, is an interval (l, u) where l and u are random values, i.e. they are not specific values but refer to all likely limits of the intervals constructed from samples of the same size. Regarding this remark, Behar [9] clarifies the difference between the interval before being computed and the interval already calculated; in the first case the values l and u are random variables that can assume particular values, and in the second occurrence, once the sample is taken, they are set and not ruled by randomness.

On the other hand, recommendations of MEN [12] for school, of Moreno & Waldegg [13] and of GAISE (Garfield *et al.*, [14]) for university, which proposed to address teaching in a more exploratory and significantly way, lead to incorporate the use of technology to enable phenomena experimentation. In statistics and probability, physical experiment simulation generates randomly data at high-speed and allows reflection on interpretations related to the frequency stability; for example, the law of large numbers evidenced by repeating an experiment many times, permits verifying that the experimental distribution approaches the theoretical [15]. In particular, it is possible to perform repetitive processes of sampling from a population data, aimed at obtaining diverse sampling information that allows to see in a better way something new that is not possible to see without technology, in order to make a conceptual re-elaboration, or in Moreno's [16] words, initiate a cognitive reorganization.

4. Confidence Level Misconceptions and Conceptual Change

Identification of misconceptions associated with the confidence interval interpretation obtained by estimating the population mean based on the sample mean and its standard error, has attracted the attention of many researchers (see e.g. Behar [9]; Cumming & Fidler [17]; Olivo [18]; Olivo & Batanero [19]; Kalinowski [20]; Yañez & Behar [21]; Cumming, Williams & Fidler [22], among others). Behar [9] recounts major misconceptions as thinking that:

- There is a 95% probability that the population mean is within the lower and upper limits of the calculated interval for a specific sample.
- 95% of the data is included in the confidence interval.
- There is a 95% probability that the confidence interval includes the sample mean.

Most of the work mentioned above reported the first misconception as the usual one. Foster [23] attributes this deficiency to the mathematics teaching in classroom when teacher gives definitions and students have no choice but to accept them without argument. Mathematical definitions, as Morgan (2005, cited in Foster [23]) says, are declarations about what is a mathematical object or idea and are a vital aspect of mathematics learning. Consequently, a strain between 'repeating of a mathematically rigorous definition' but not well internalized, and expressing it imperfectly but reflecting possible hints of comprehension, arises, and it is necessary to establish whether the center of the student assessment is to determine the procedural fluency in a mathematical technic or in a very different way, try to discover the underlying understanding in their productions [23]. This tension creates a duality that contrasts what Skemp [24] calls "instrumental understanding" and "relational understanding".

Similar to the emergence of mathematics in history, learning can be seen as an evolutionary process, in which students build their knowledge gradually founded in their effort and errors. In particular, the learning theory called conceptual change, grounded in research in science education but built on Piaget ideas, constructivism and situated learning theories posits that students learn when their existing conceptions are challenged; hence, the research on 'conceptual change' is also a research on 'misconceptions'. In delMas, Garfield & Chance [25] (p. 5) words, "students who have misconceptions or misunderstandings need to experience an anomaly, or contradictory evidence, before they will change their current conceptions".

Conceptual change is a process that happens over time and there is not necessarily an instantaneous reorganization and replacement of concepts [26]. This process, Gunstone [27] and other researchers remark, "involves the learner recognizing his/her ideas and beliefs, evaluating these ideas and beliefs (preferably in terms of what is to be learned and how this is to be learned), and then personally deciding whether or not to reconstruct these existing ideas and beliefs"; in summary, conceptual change occurs when students: "recognize, evaluate, decide whether to reconstruct" [27].

The "shift or restructuring of existing knowledge and beliefs is what distinguishes conceptual change from other types of learning; learning for conceptual change is not merely accumulating new facts or learning a new skill; in conceptual change, an existing conception is fundamentally changed or even replaced" [28].

According to Vosniadou [29], instruction for conceptual change of students' "naive theory of how something works based on previous experience", should create a perturbation in their minds by addressing an "inconsistent with their existing mental representations". As Davis [28] states, teaching for conceptual change primarily involves 1) uncovering students' preconceptions about a particular topic or phenomenon and 2) using various techniques to help students change their conceptual framework".

5. Class Activities

The general idea behind the activities design is that students will be concerned about their existing knowledge and see the need for changing it. Therefore, the work proposed intent to question students' conception of confidence level so they can conclude that there is an anomaly, and then based on simulation work, make a conceptual change of that notion.

The planned activities were implemented along several consecutive class sessions, with college students who had already taken a course in statistical methods. The design of the instruction followed the guidelines suggested by the research design methodology, such as hypotheses formulation about how students' learning is expected to evolve throughout the instruction. The planned activities are divided into seven parts, which will be briefly overviewed in this paper. Afterwards, the focus will be on the third, fourth, sixth and seventh parts, which are the ones we are interested in discussing, and the learning assumptions that underlie them.

The context for the class activities suggests a problem of measurement and statistical characterization of the IQs of a thousand school boys and girls, who make up the target population. In the first part the students work around the difference between the notions of population and sample of individuals, as well as around data sets related to the context of the situation which describe a trait of the individuals. The second part revises random sampling of the given population and sample mean computation of IQs; this part aims at clarifying the difference between parameters and statistics. The third part addresses the interpretation of confidence level and questions its meaning in order to contribute to its re-conceptualization. The fourth part proposes a manual simulation that pursuits help students to note the frequency idea and to initiate the approach to a new interpretation of the confidence level, the frequentist one. The fifth part encourages working in groups for revisiting and comparing the interpretations enunciated in the previous work and in manual simulation. In the sixth part, we appeal to computer simulation with Excel to help students to explicitly realize, verify and strengthen the frequentist interpretation of confidence level as well as gain confidence in their own findings. Finally, in the last part we put forth a typical textbook situation so that the student would interpret confidence level based on the gained ideas.

The baseline scenario for the proposed work, in the third and subsequent parts of the class activities, is that the students' usual interpretation of the confidence level, once the interval is built, coincides with the misconception described earlier, i.e. that the confidence level is the probability that the population mean is within the lower and upper limits of the calculated interval. In other words, we assume that students consider that probability implicitly refers to the confidence interval in itself as an event of the sample space, understanding the latter as the real number line.

Students start the work, in this part, by interpreting the confidence level before the calculation of the confidence interval. We believe that the frequentist approximation of the confidence level may be more visible to students at a time prior to the construction of the confidence interval, when the probability that constitutes the confidence level is established. Since there are no numbers yet to determine the interval, it is possible that students' statements will be close to a frequentist interpretation of probability, and do not refer to the specific interval limits nor imagine the real number line as the sample space or the reference set.

Next, we look forward to confront students' failed idea of confidence level as they work with two random samples of the same size, generating two non-overlapping confidence intervals. Then, students should make the interpretation of the two confidence levels, which we anticipate would fit the mentioned misconception: confidence level is the probability, 90% in this situation, that the population mean is contained in the built interval, being identical for both intervals. Naturally students will see that the confidence intervals calculated do not overlap, and we expect that, due to this fact, to the consideration of the interval and its complement in the real number line, to their knowledge of probability, as a number between 0 and 1, and of complementary probability, students will recognize the conflict in their interpretations of the confidence level applied to both intervals, since such interpretations contradict the probability rule for disjoint events and the probability rule of monotonicity ($P[E] \leq P[F]$ if $E \subset F$. Hence, we aspire that students note that there is something unsuitable about their confidence level conception and they will question themselves about it.

Third part. Working with two samples (work in pairs)

1. If you were to estimate the IQ mean of the population, by using a 90% confidence interval, describe your interpretation of the confidence level, before computing the interval.
2. For the next sample* of twenty students' IQs, construct a 90% confidence interval for the population mean. Refer to it as I_1.

   149 129 119 130 97 128 129 107 98 122
   136 113 115 117 118 142 137 120 134 140
3. Describe your interpretation of the confidence level of this interval.
4. For the next sample of twenty students' IQs, also obtained randomly and which is also thought as representative of the population data, construct a 90% confidence interval for the population mean. Refer to it as I_2.

   101 107 130 101 115 104 91 91 121 109
   104 125 113 98 110 119 102 92 111 120
5. Describe your interpretation of the confidence level of this interval.
6. Explain whether there is an inconsistency with the confidence level interpretations for the two built intervals (I_1 and I_2).

* The sample was obtained randomly and it is considered representative of the population data.

Since it is quite possible that students do not detect the inconsistency in the interpretation of the confidence level and insist in the referred misconception, the work is then focused on representing both intervals on the real number line, expecting that this concrete and familiar mathematical structure will help to make the conflict more noticeable. In the first place, students will see that one interval is a subset of the complement of the other, and vice versa. In the second place, students should note -founded on their knowledge of probability- that the maximum probability of the complement of the interval is 10% (since the confidence level is 90%), and that the probability of the event, as interpreted by them, related to the other interval, that is a subset of such complement, was established as 90%. Finally, students should recognize that this is absurd, as it contradicts the rule of monotonicity which establishes that if $E \subset F$ then $P[E] \leq P[F]$, or be aware that the previously considered probabilities, related to the intervals and their complements, must satisfy the probability rule of the union of disjoint events, which in this case, will produce a probability greater than 1. This way we expect that students will find out the conflict and realize the presence of something inappropriate in their idea of confidence level.

7. For the intervals built before, do the following:
   a. Plot the two intervals in the same real number line.
   b. Find out whether the intervals overlap. Identify the complement of each interval.
   c. Explicitly express the relationships between an interval and the complement of the other.
   d. Complete the following table by noting the intervals I_1 and I_2, the confidence level interpretations made on items 3 and 5 for these intervals, and the interpretation of the probabilities of the intervals complements, $I_1^c$and I_2^c.

Confidence interval	90% Confidence level interpretation	Interpretation of 10% associated to the interval complement
I_1		
I_2		

8. Considering the subset relationships found in item 7c, describe the relationship between the interpretation given to the 90% confidence level associated with the interval I_1 and the interpretation given to the 10% associated with the complement of the interval I_2.
9. Considering the subset relationships found in item 7c, describe the relationship between the interpretation given to the 90% confidence level associated with the interval I_2 and the interpretation given to the 10% associated with the complement of the interval I_1.
10. Based on the above, explain whether there is an inconsistency with the confidence level interpretations for the two built intervals (I_1 and I_2).
11. So finally, what is your pronouncement about the interpretation of the confidence level associated with the intervals (I_1 and I_2)?

In order to allow students to reassure their finding about an irregularity in the confidence level interpretation, afterwards, we orally interact with students upon a prepared script depending on possible responses given to the table presented on item 7d. For example, the next script is used in case the students' responses would be the mentioned misconception.

Possible students' responses and interaction

Confidence interval	90% Confidence level interpretation	Interpretation of 10% associated to the interval complement
$I_1 = (118,79; 129,20)$	*The population mean is in I_1 with a probability of 90%*	*The population mean is in I_1^c with a probability of 10%*
$I_2 = (103,88; 112,11)$	*The population mean is in I_2 with a probability of 90%*	*The population mean is in I_2^c with a probability of 10%*

The teacher questions the students: What are the events associated with the declared probabilities?
The teacher asks about the probability rule of an event contained in another event, and requests them to express the relationship between the probabilities of those events.
The teacher raises the question: What could then be the probability that the population mean is in the second interval obtained?
Then the teacher asks if the above relationships are consistent.

After manual collecting of random samples and building the correspondent confidence intervals, we encourage students to compare likeness between the percentage of intervals capturing the population mean and the confidence level established, to exhort once again, the frequentist interpretation of confidence level.

Fourth part. Genrating data for manual simulation

12. Collect a new random sample of 20 IQs of the population. For this purpose, generate 20 random numbers from 1 to 1000, and according to the provided list, identify the respective IQs of students. Find the sample mean and deviation.
13. With the collected sample, calculate three, 90%, 95% and 99%, confidence intervals.
14. Collect four more random samples and repeat the above process for each sample.
15. Gather the information from your peers to complete the table below.

Number of 90% intervals containing μ	Number of 95% intervals containing μ	Number of 99% intervals containing μ
Percentage of 90% intervals containing μ	Percentage of 95% intervals containing μ	Percentage of 99% intervals containing μ

16. Compare each percentage in the table with the respective confidence level established when the intervals where built. Discuss what you observe.
17. Given the relationship hinted at in the table, describe what might be the interpretation of the confidence level.

We then presume that through computer simulation oriented to collect 100 samples and build the respective 95% confidence intervals, students recognize conceptual elements related to confidence intervals, as suggested by delMas, Garfield & Chance [25], and verify that about 95 of the 100 calculated intervals capture the population mean, corroborating the frequentist interpretation. Finally, we hope this evidence helps them deciding to reconstruct their current conception.

Sixth part. Simulation with Excel

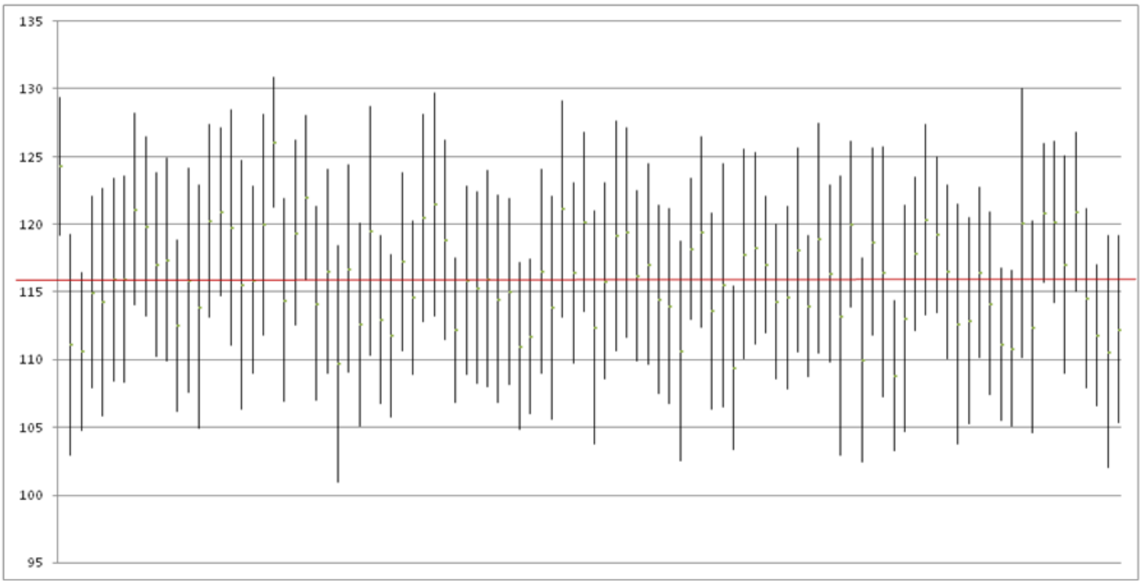
26. Identify the elements involved in the computerized procedure of simulation.
27. Perform the simulation and describe the procedure to simulate sample selection and intervals construction.
28. Complete the following table, based on the simulation.

Simulation	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Number of 90% intervals containing μ										
Percentage of 90% intervals containing μ										

29. Compare each percentage in the table with the respective confidence level established when the intervals where built. Discuss what you observe.
30. Given the relationship observed, what do you think would happen between the percentage of intervals that capture the population mean and the confidence level established to build the intervals, if you repeat the simulation 500, 700 or 1000 times?
31. According to the answer in the previous item, and to the whole work developed, propose an interpretation for the confidence level that can be applied to a collective of confidence intervals and not to a specific interval, so incompatibilities are not generated.

Simulation example

The figure shows a simulation of confidence intervals for the population mean assuming that the population deviation is unknown. The parameter to be estimated is the population mean, usually unknown, but set for the simulation and displayed by the highlighted line. The simulation renders random collection of 100 samples of size n = 20 and the respective 95% confidence intervals, represented by the vertical segments. Although the samples are of the same size, intervals in the graph does not always have equal lengths, since for each sample, standard deviation σ was estimated based on the sample deviation *s*, which can vary from sample to sample.



6. Students' Work around the Interpretation of Confidence Level

In this section we present students' work in some of the class activities, trough responses and discussions of each case study, named as stated before, **C**, **A**, and **B**. Also, we exhibit statistic graphs showing trends in confidence level interpretations exposed by the entire student group, firstly, when building the confidence intervals and secondly, while simulations development.

A priori interpretation of confidence level, i.e. before building the intervals

Students' confidence level interpretations before calculating confidence intervals are the following.

> **C**: *The confidence level tells us that if we took several samples and calculate the mean, 90% of the samples, or better, about 90% of the samples contains the mean in the confidence interval*
>
> **A**: *An interpretation to what the confidence level means, is the probability that the mean belongs to the interval*
>
> **B**: *We can interpret the confidence level as the distance between the interval's lower and upper limits, the greater the confidence level, the smaller the distance between the two limits and greater reliability*

C's interpretation suggests an idea of frequency although it refers to samples rather than intervals; at the end comprises confidence interval in singular, alluding to a specific interval or perhaps wanting to refer to each of the respective intervals obtained from the samples. Here, we begin to evidence the students' difficulty to precise in words what they really understand about confidence level. **A**'s interpretation coincides strictly with the misinterpretation that was behind class activities. **B** identifies the confidence level with the interval length and refers to greater reliability when the interval is smaller, equally a misconception reported in the literature (see e.g. Behar [30], [9]). The fact of referring to the relationship between confidence level and interval length, could be originated in the usual teaching emphasis on it.

Confidence level interpretation in front the conflict

Students' confidence level interpretations after constructing the two disjoint intervals are discussed.

Case study	Confidence interval	90% Confidence level interpretation	Interpretation of 10% associated to the interval complement
C	I_2 *(104,112)*	*With a 90% confidence level the population mean is in the interval*	*With a 10% confidence level the population mean does not belong to this interval*
	I_1 *(118.6, 129.4)*	*With a 90% confidence level the population mean is in the interval*	*With a 10% confidence level the population mean does not belong to this interval*
A	I_1 *(103.93,112.47)*	*The probability that the mean is contained in the interval I_1 is 0.9*	*The probability that the mean is contained in I_1^C is 10% or even less*
	I_2 *(118.66,129.33)*	*The 90% confidence level refers that the population mean is contained in I_2*	*The probability that the mean is contained in I_2^C is 10%*
B	I_1 $105,24 \leq \mu \leq 132,75$	*There is a reliability of 90% that the mean is in the interval I_1, taking into account the given data*	*There is a reliability of 10% that the population mean is in the interval I_1^C. This being much larger than I_1*
	I_2 $101,20 \leq \mu \leq 115,20$	*There is a reliability of 90% that the mean is in this interval I_2, taking into account the given data*	*There is a reliability of 10% that the population mean is in the interval I_2^C this being much larger than I_2*

C's, **A**'s and **B**'s interpretations are consistent with the confidence level misinterpretation assumed for class activities. In spite of expressing the same confidence level interpretation for the two disjoint confidence intervals, students do not notice the conflict; it seems that in this moment, they consider the probabilities regardless of the underlying set relationships of the intervals; neither, they display thinking of sample spaces associated with the stated probabilities. When **B** emphasizes that the interval's complement is larger than the interval, exists the possibility that she might be seeing the monotonicity rule of probability.

When asked about the conflict between interpretations, students add,

> **C**: *No, because in each response a significance level is given, which says that an error margin exists, in this case a big one (10%)*
>
> **A**: *The confidence level regarding I_1 is the probability to find the population mean, this probability is 0.9, which also implies that there is a 0.1 probability that the population mean is not in the built interval (I_1). The 90% confidence level refers to the probability that the population mean μ is in I_2, so in this way it is assigned a high probability of finding this measure in the built interval. There is no incompatibility in the interpretations, since the two refer to the probability that the population mean is found in a built interval from a given sample*
>
> **B**: *Yes, since I_1 and I_2, where we are placing the population mean are varying according to the produced samples*

C and **A** still do not see the conflict. **C** alludes to a "significance level" and "error margin" that seems to be the confidence level and complementary percent, respectively; besides, the reference to "error margin" suggests that regardless of the probabilistic quantification none of the interpretations is 100% secure, so they can be true or not, and either case is valid; thus the possible incompatibility between them is eliminated. **A** explains the meaning of the complementary probability and refers to the fact that there are two samples and two correspondent intervals, and in a similar way the fact that none of the interpretations is 100% secure, eliminates the potential incompatibility. **B** answers the question about an incompatibility, saying "yes" but linking it to sample variation. It is worth to note that they continue to ignore the necessity of thinking of sample spaces for the probabilities.

Later on, students say:

> **C**: *An interval is found with a 90% probability, contained in an interval of 10% probability with independent events and we have a contradiction with: if a $\subset$ b then Probability of a < Probability of b*
>
> **A**: *Since $I_1 \subset I_2^C$ then the probability that the mean is in I_2^C is 10% or even less. Since $I_2 \subset I_1^C$ then the probability that the mean is in I_1^C is 10% or even less. In the I_1 interval, the probability that the mean is contained here is 90%, and the probability that the mean is contained in the I_2 interval is 90% also, then when observing them as two intervals of the same population, the incompatibility in I_1 and I_2 is found in terms of their probability, because it exceeds 100%*
>
> **B**: *Subset relation $I_1 \subset I_2^C$. It is assumed that the probability that the mean is inside I_1 is 0.9 yet at the same time is 0.1 given the subset relation. Subset relation $I_2 \subset I_1^C$. There is an incompatibility between the interpretations since they contradict the probability theorem, $A \subset B \rightarrow P(A) \leq P(B)$. Where $A = I_1$ $B = I_2^{C.}$*

Here **C**, **A** and **B** recognize the numerical contradiction between probabilities when applying the probability rule of monotonicity (P [E] <P [F] if E ⊂ F) and the probability rule for disjoint events, and therefore they realize the incompatibility between the interpretations.

Student's responses to new questions about possible conflict between interpretations, are:

> **C**: *Concerning our interpretations, we do not find incompatibilities, but the fact that there are incongruities in the analysis done makes us think we do have incompatibility with the interpretation that should be given to the confidence level* […] *we still have doubts that we have not been able to interpret them properly. After answering the previous questions we reflect and maintain concerns* […] *because they have reported inconsistencies*
>
> **A**: *The confidence level indicates the probability that the population mean is contained in an interval, however it generates ambiguity when relating two intervals since it can be thought that such probability exceeds 100%*
>
> **B**: *There is an error in the interpretations, and it is necessary clarifying to what we refer with the confidence level of the problem*

The ambivalence in **C**'s and **A**'s conception is manifested. They insist in the confidence level interpretation assumed for the class activities, albeit the contradiction found, and at the same time hesitate about the interpretations' appropriateness. **C** again admits his concern and it seems that he starts to question his confidence level conception. **B** recognizes also an error in the interpretations, but similar to **A**, particularize the conflict to the situation with two disjoint intervals. We sense a kind of subjective reasoning in the way students perceive something different about the probability after gathering information originated by the detected conflict.

Confidence level interpretation while simulations occur

Now we discuss the effects of simulations in students' interpretations, firstly, through the cases studied, and then within the whole group of students.

> **C**: *The confidence level refers to the probability of μ being contained in a certain interval*
>
> **A**: *The confidence level is interpreted as the probability that the mean μ is contained in the constructed interval and the conclusion is that the greater the confidence level the greater the percentage of intervals that contains μ*
>
> **B**: *The greater the confidence level the greater the probability of finding the population mean in that interval*

Despite the recognition of a contradiction, **C** and **A** go back to the misinterpretation expressed formerly, but on account of the reference made by **A** to the percentage of intervals, we dare to say that his interpretation is changing since now includes an approach to the idea of frequency. **B** mentions the relation found between the confidence level and the chance to encounter the population mean in the interval.

While working with computerized simulation, students say,

> **C**: […] *If we make 100* [intervals] *maybe there is one in which* [the population mean] *there will not be, in 99 yes, but in one, does not.* [The confidence level] *is the frequency of appearance of μ in many intervals, or must be something similar* […] *It is that looking at the confidence level locally in one interval... no. Maybe in the last class exercise it does not make sense to look at the confidence level for one interval, because for 90%* [μ] *will not fall in the examples* [the disjoint intervals]. *There are some that yes and some that no. There would be to look with many intervals. The confidence level refers to the frequency of the population mean contained in certain amount of different intervals* […] *indicates roughly the percentage of intervals that contain the population mean when we choose different samples randomly*
>
> **A**: *We talked that the confidence level was given only in that interval, i.e. that there was a probability that the mean was in that interval. Yet we do not relation as such any more* […] *The first interpretation that we did for the confidence level was for one interval, now this interpretation that the table is showing us* […] *what we saw is that most intervals intersect, then the confidence level does not go in the interval as such but in comparing different intervals*
>
> **B**: *With a 99% confidence interval the probability that the population mean is in that interval is 100%, greater than* [the probability of the] *90% and 95% confidence intervals. We think that the percentage of 90% intervals that contain μ is even nearer to the managed confidence level. The confidence level of a group of intervals is most accurate when the group of intervals is larger, so the confidence level can be interpreted as the percentage of intervals of specified percentage (90, 95, ...), containing μ*

Simulation tasks make possible that students undergo in practice, and also confirm, the idea of several intervals containing the population mean; even more, as result, all of them expose a frequentist interpretation for the confidence level or at least an approximation to it.

Likewise, the simulation work arouses interrogations in students about the possible meaning of an estimation, building just one confidence interval, so they declare that an interpretation connected to the idea of frequency makes more sense than tied to a single interval.

The following graph shows similar outcomes for the entire group. Before computer simulation, while students worked with manual simulation, a salient number of responses evidences that students noted that varying the confidence level implies that the percentage of intervals that contains the population mean, varies directly; so this appears to be a start point to the idea of frequency. Then, during simulation with Excel, there are more precise explications about connection between confidence level and percentage of intervals comprising the population mean, so responses coincident with, or close to, a frequentist interpretation of confidence level, are prominent.

Confidence level interpretation after solving a textbook problem

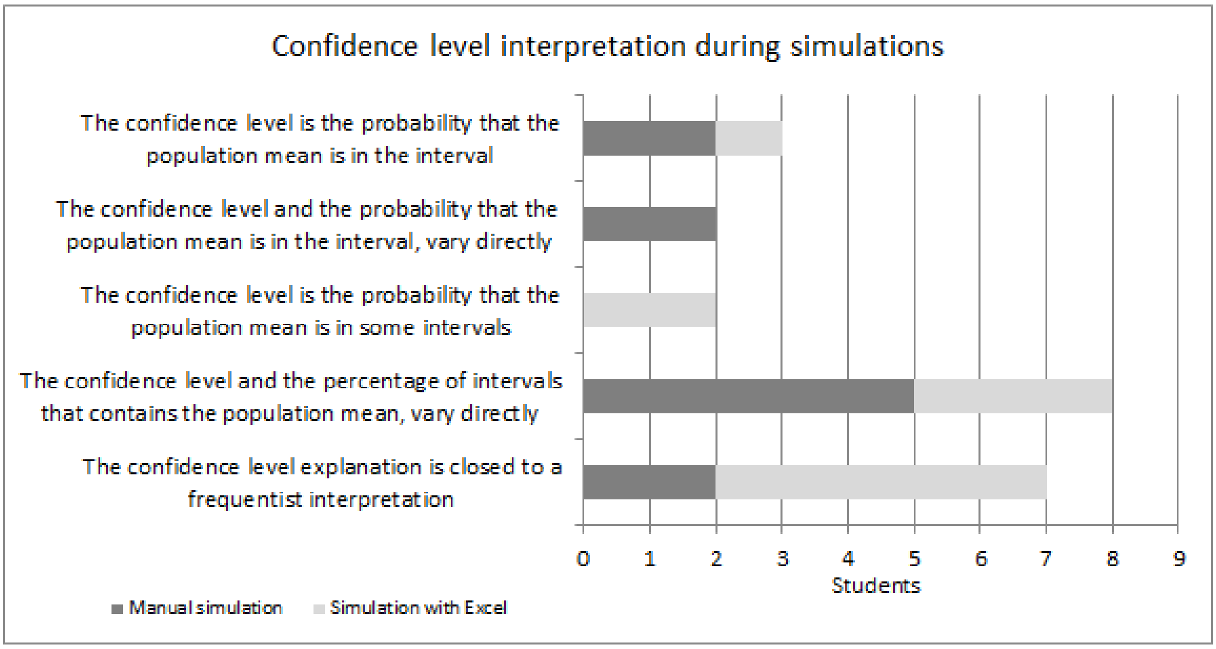
After the development of class activities, students' confidence level interpretations in the solution of a typical textbook estimation problem, by calculating a confidence interval, are shown.

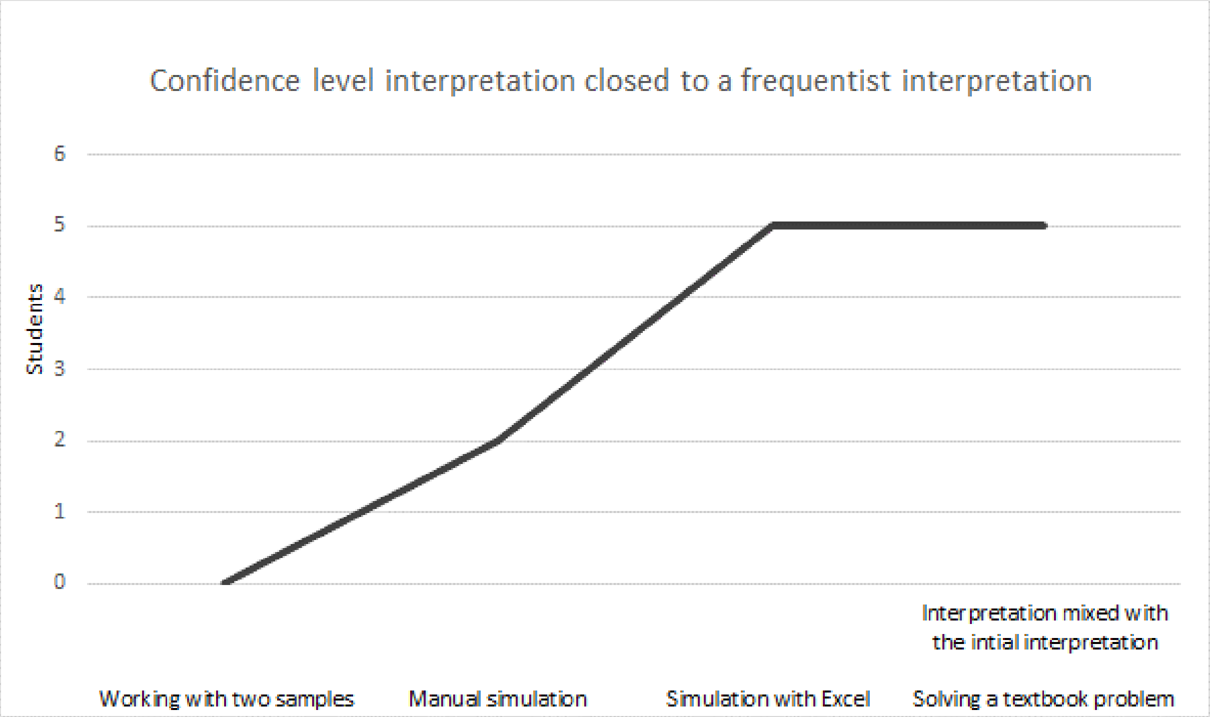
> **C**: *In this confidence interval it can be assert that in 90% of the samples of 6 liquid measures, the population mean will be in this interval*

> **A**: *The probability that μ □is in the interval (4.8183, 4.9815). What we think is that there is a 90% confidence that the true liquid conductivity they are selling, is in the interval (4.8183, 4.9815)*

> **B**: *The obtained confidence interval shows that with a 90% confidence level the alleged conductivity may be 5 (very close to the obtained sample mean that is in the interval). This confidence interval determines that with 90% the population mean is in this interval*

Despite the conflict noticed before and the expressions, close to a frequentist interpretation, for the confidence level, students return to their initial misinterpretation, maybe due to the type of situation, usually solved in that way. Although **C** evokes the idea of frequency when referring to "90% of the samples", contradicts it pointing to the specific interval. **A** expresses again and reaffirms the alleged interpretation for class activities in the context of the situation. It is possible that when **B** uses the word "may", sees the calculated interval as one that contains the media, and consequently approach to an interpretation associated with frequency; however, in the last part of her answer refers again to the assumed interpretation for class activities.





The raised questioning and the work with simulations helped students to see an anomaly in their conception and provoked to speak of frequency but do not appear to have created a conceptual change in students, as revealed in the case studies' previous responses and in the whole group replies shown in the above graph. Maybe it is necessary more time and alternative strategies to instigate students deciding to reconstruct their idea. Neither the use of technology reflects a cognitive reorganization in the sense indicated by Moreno [16]. Although the three students considered as case studies recognize the inconsistency between the confidence levels as probabilities, and thus show their knowledge of formal probability rules, the probability notion manifested does not attend to identify a sample space; this does not help them to really see what is the problem with assigning the same confidence level interpretation for two disjoint intervals or to stick to the envisioned interpretation related to frequency.

7. Conclusions

The effect of the designed activities oriented to challenge the students' conceptions appears to be not perdurable. As seen before, students show confidence level interpretations that fluctuate, i.e. sometimes during the development of the activities their interpretations are close to the frequentist interpretation, but at other moments, their expressions match the ones they indicated previously, reported as inadequate. The work generates the expected questioning about their confidence level conception and they express concern regarding the role of the confidence level in the estimation process, though it is clear that the intended perturbation was not strong enough in the students for a conceptual change of their existing confidence level conception. Despite that students see the contradiction generated by the probability rule for disjoint events and the probability rule of monotonicity, initially they overlooked it and solve the incongruity, explaining probability as a statement not hundred percent trustworthy.

The relentless presence of the misconception about the notion of confidence level, reveals some points to be considered when the instruction of confidence intervals takes place. To begin with, the possible meaning of the estimation process based on a single sample and therefore on the construction of a single confidence interval. The perplexity expressed by the students regarding this is understandable since once the interval is computed, stating that the confidence level determines the number of intervals that contain the parameter, does not seem to shed much light upon the inference process being done. Consequently, the construction of only one confidence interval appears as a poorly accurate estimate, even as a useless estimate; of this particular interval we can only say that it may or may not capture the parameter, i.e. it may, or may not, be one of the percentage of possible intervals built, containing the parameter.

The reply of several experts to this doubt is the endorsement of the method for constructing intervals in the process of estimation, already mentioned, reminded by Behar [9] when he says "you never know" if the confidence interval constructed captures the parameter but "based on the credentials of the procedure you can act as if the particular interval had caught the real mean, with the risk associated to the generator process". Nonetheless, this approach is far from obvious for the students, just as Behar [30] has shown in a study where students do not associate confidence to an interval generator random mechanism.

In search of making sense and perceiving a practical utility of a confidence interval, instruction could propose situations where the built interval helps in decision making with uncertainty, e.g. when a desirable mean has been set, a confidence interval allows to conclude depending on whether the interval contains the mean or not. Also, situations where computing a confidence interval is helpful in hypothesis tests; for example, when conjecturing about the parameter before building the confidence interval, computing it and checking whether the parameter belongs to the interval allow to reject or validate the conjecture.

Another crucial point is the language used by teachers and most textbooks that does not seem to infuse sense to this estimation process. Language rarely refers explicitly to the percentage or number of possible intervals to be built, and distinctions in the ways of expressing the estimation are subtle, if there are any. For example, Moore [31] statement that "the confidence level gives the probability that the interval will capture the true parameter value in repeated samples" is a shorthand for "we got these numbers using a method that gives correct results 95% of the time". Montgomery & Runger [10] as well, point out that the statement "the interval $[l, u]$ captures the true value of μ with confidence 100× (1 - α)%", where l and u are specific values, has a frequentist interpretation because despite not knowing whether this assertion is true for this specific sample or not, you have the assurance that the method produces trustworthy statements 100×(1- α)% times. However, students are limited to the words in the statements that can be understood as 'μ is in (l, u]) with probability 100×(1- α)%', expression not adequate as mentioned previously.

Robinson-Cox (1999, cited in Foster [23]) argues that the essence of the difficulties is to place the confidence in the calculated interval rather than on the process by which it is determined; for example, he notes that "students mistakenly say 'The probability that μ is in (7.5, 9.2), the calculated interval, is 0.90' instead of saying 'The process by which the interval (7.5, 9.2) is calculated, includes μ 90% of the times'" (p. 81). Most textbooks do not help much in this situation, since they register the term 'probability' or its synonyms linked indistinctly to the two objects involved in the confidence level interpretation: to the interval, usually when it is defined, or to the parameter, at the conclusions of the proposed tasks. For example, Moore [31] remarks that "the confidence level gives the probability that the interval will capture the true parameter value in repeated samples", i.e. the probability is tied to the interval, but also says "we are 95% confident that the unknown μ lies between 26.2 and 27.4 [interval limits]", connecting the confidence level to the parameter. It can be seen then that the frequentist interpretation of the confidence level is approached lightly, and expressions that might be conflicting and confusing are accepted as equivalent. We claim that the frequentist interpretation would manifest most clearly for students, if the term 'probability' or its synonyms are explicitly associated with the estimation process or even with the confidence interval, but no with the parameter.

Besides, even though in textbooks the term 'probability' is used to define the confidence level, for example when Ross [32] establishes the confidence level as "the probability that the interval contains the parameter", the estimation tasks illustrated there refer rather to "degree of confidence", "confidence" or "security". For instance, Moore [31] expresses "…we are 95% confident…", Montgomery & Runger [10] concluded "...is the interval of fair values for the mean with 95% confidence", Christensen [33] declares "we are sure that 95% of the confidence intervals…". It seems then that the word 'probability' is awarded a stronger connotation, linked to the measurement of the occurrence of an event and thereby it is not used in the practice. Is in this sense that Christensen [33] argues that θ may or not be in the computed confidence interval, for this reason he talks in terms of reliability instead of probability, and that Montgomery & Runger [10] admit that, since one confidence interval is constructed in the practice only, which may include or not the real value of the estimated statistic, "it is not reasonable to assign a probability level to this specific event".

The problem to eradicate the misconception and to gain clarity on the interpretation of the confident level linked to the idea of frequency, can also be due to students' difficulties conceptualizing probability notion and with the diversity of ways in which probabilities can be assigned, so much, that students' conception of probability appears to agree with some of the conceptions that researchers as Batanero [34] [3] have proposed. In the development of class activities, students conforming the case studies conceive probability as a numerical value, without necessarily being tied to a specific set of reference, i.e., to a sample space; even after passing through at least one course of probability, they do not show concern or need to identify the sample space in their responses regarding the confidence level as probability. Neither they appeal to Laplace's probability conception stressed in school, in order to anyhow explore or verify, the conflict detected. Thereby, we glimpse a basic understanding of probability that recalls the use of primary intuitions, in the spirit revealed by Fischbein [35], which is also related to difficulties in identifying and restricting the sample space, according to Totohasina [36]. This conception can be associated to the intuitive interpretation of probability which is connected to naive knowledge, opinions and beliefs, and the use of colloquial phrases to express it. In students' formulations it is also possible to perceive an incipient reasoning that approaches Bayesian subjective probability

[3] Batanero [34] pointed out different interpretations of probability that are accepted today: the intuitive, the Laplace's, the frequentist, the subjective and the axiomatic.

conception[4], when students consider new evidence to weigh the reasonableness of the probabilities stated.

The fact that students realize the infraction of basic laws of probability, but continue to accept the alleged confidence interval interpretation as valid, turns on the alarms in two directions. On the one hand, it questions the teaching practices, seeing that students' conception of probability remains in an intuitive state despite the instruction; on the other hand, from a conceptual point of view, it warns us about the need to reassess the sense of working with confidence intervals and the confidence level frequentist interpretation as the trustworthy interpretation; inasmuch as Bayesian perspective of probability would seem promising, it also encourages addressing confidence intervals from this angle.

The final point is related to the limited study of variability in statistics courses, firstly, because statistics class work deals merely with variation of data sets; secondly, because statistical inferences based on 'induction' from a single sample, seems to truncate the possibility of thinking in the variation between samples. These reasons help to position the interpretation of confidence level as the probability with the sample space constituted by the population from which the sample is taken, and not formed by the set of all possible samples of a given size that can be taken from the population. Besides, students are not aware that thinking that the parameter will be in the specific built interval 95 of 100 times, is equivalent to thinking that the parameter sometimes is in the interval and other times it is not, and therefore the parameter will vary whereas it is unique; it is evident that students do not approach to seeing that variation is in the possible intervals to construct. So according to Cumming & Fidler [17], students visualize confident intervals as descriptive statistics and ignore their inferential nature.

Hence, to familiarize students with the frequentist interpretation, introductory descriptive statistics courses should include tasks that make variation perceptible, especially, variation between the values associated to estimators generated from different samples of the same size, i.e. tasks that allow students to account for the variation that is present in the estimations linked to the variation of the corresponding samples.

REFERENCES

[1] Andrade, L., Fernández, F. & Álvarez, I. (2014). Fostering changes in confidence intervals interpretation. In K. Makar, B. de Sousa & R. Gould (Eds.), Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9), Flagstaff, Arizona. http://icots.info/9/proceedings/pdfs/ICOTS9_C187_ALVAREZ.pdf

[2] Lesh, R. & Kelly, A. (2000). Multitiered teaching experiments. In A. Kelly & R. Lesh (Eds.), Handbook of research design mathematics and science education (pp. 197-230). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

[3] Steffe, L. & Thompson, P. (2000). Teaching experiments methodology: Underlying principles and essential elements. In A. Kelly & R. Lesh (Eds.), Handbook of research design in mathematics and science education (pp. 267–306). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

[4] Cobb, P. (2000). Conducting teaching experiments in collaboration with teachers. In A. Kelly & R. Lesh (Eds.), Handbook of research design in mathematics and science education (pp. 307–333). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

[5] Neiman, G. & Quaranta, G. (2006). Los estudios de caso en la investigación sociológica. In I. Vasilachis (Coord.), Estrategias de investigación cualitativa (pp. 213-234). Barcelona: Editorial Gedisa, S.A.

[6] Ponte, J.P. (2006). Estudio de caso en Educación Matemática. BOLEMA, 25, 103-132.

[7] Colás, M. & Buendía, L. (1992). Investigación educativa. Sevilla: Ediciones Alfar.

[8] Davies, H. (1998). What are confidence intervals? http://www.evidence-based-medicine.co.uk

[9] Behar, R. (2007). ¿Estamos buscando el ahogado aguas arriba? El caso de la estimación con intervalos de confianza. Primer Encuentro Nacional de Educación Estadística (ENAES), Bogotá. http://www.encoedest.org/C03_Behar.pdf

[10] Montgomery, D. & Runger, G. (2004). Probabilidad y estadística aplicadas a la ingeniería. México: Limusa, S.A.

[11] Freund, J. & Walpole, R. (1990). Estadística matemática con aplicaciones. México: Prentice-Hall Hispanoamericana.

[12] MEN (1998). Lineamientos curriculares de matemáticas. Bogotá, Colombia: Cooperativa Editorial Magisterio.

[13] Moreno, L. & Waldegg, G. (2002). Fundamentación cognitiva del currículo de matemáticas. In Memorias Seminario Nacional de Formación de Docentes: uso de nuevas tecnologías en el aula de matemáticas. Bogotá: MEN.

[14] Garfield, J., Aliaga, M., Cobb, G., Cuff, C., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P. & Witmer, J. (2005). Guidelines for assessment and instruction in statistics education (GAISE): College Report. Alexandria, VA: American Statistical Association. http://www.amstat.org/education/gaise

[15] Drier, H. (2000). Children's probabilistic reasoning with a computer microworld (Doctoral thesis). Charlottesville: University of Virginia.

[16] Moreno, L. (2002). Calculadoras algebraicas y aprendizaje de las matemáticas. In Memorias Seminario Nacional de

[4] According to Batanero [34], the assignment of subjective probability is based on transforming established probabilities before an experiment, in probabilities that include information from the observed data after performing the experiment, trough Bayes theorem The established probability depends on the knowledge and experience of the person who assigns it, and is always conditioned by his/her system of knowledge so it may differ for different people. In this prospect, repeating the experiment under the same conditions it is not required anymore. This conception widens the field of application, in particular to the study of decisions in economy, diagnostic and other. Currently, Bayesian probability conception is applying to all types of uncertain events, although the controversy over the scientific study of subjective probabilities, continues.

Formación de Docentes: uso de nuevas tecnologías en el aula de matemáticas (pp. 93-98). Bogotá: MEN.

[17] Cumming, G. & Fidler, F. (2005). Interval estimates for statistical communication: problems and possible solutions. IASE/ISI Satellite.http://IASE-web.org/documents/papers/sat2005/cumming.pdf

[18] Olivo, E. (2008). Significado de los intervalos de confianza para los estudiantes de ingeniería en México (Doctoral thesis). Granada, España: Universidad de Granada.

[19] Olivo, E. & Batanero, C. (2007). Un estudio exploratorio de dificultades de comprensión del intervalo de confianza. Unión, 12, 37-51. http://www.mty.itesm.mx/dtie/deptos/m/ma00-835/Articulos-nuestros/Olivo-Batanero(2007)Un_estudio_exploratorio_Intervalos.pdf

[20] Kalinowski, P. (2010). Identifying misconceptions about confidence intervals. In C. Reading (Ed.), Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8), Ljubljana, Slovenia. http://icots.net/8/cd/pdfs/contributed/ICOTS8 C104_KALINOWSKI.pdf

[21] Yañez, G. & Behar, R. (2010). The confidence intervals: a difficult matter, even for experts. In C. Reading (Ed.), Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8), Ljubljana, Slovenia.

[22] Cumming, G., Williams, J. & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. Understanding Statistics, 3, 299-311.

[23] Foster, C. (2014). Confidence trick: The interpretation of confidence intervals. Canadian Journal of Science, Mathematics and Technology Education, 14(1), 23-34. http://www.tandfonline.com/doi/pdf/10.1080/14926156.2014.874615

[24] Skemp, R. (1976). Relational understanding and instrumental understanding. Mathematics Teaching, 77, 20–26.

[25] delMas, R., Garfield, J. & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. Journal of Statistics Education, 7. http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm

[26] Mayer, R. (2002). Understanding conceptual change: A commentary. In M. Limon & L. Mason (Eds.), Reconsidering conceptual change: Issues in theory and practice (pp. 101-114). Dordrecht, The Netherlands: Kluwer Academic Publishers.

[27] Gunstone, R. (1994). The importance of specific science content in the enhancement of metacognition. In P. Fensham, R. Gunstone & R. White (Eds.), The content of science: A constructivist approach to its teaching and learning (pp. 131-146). London: The Falmer Press.

[28] Davis, J. (2001). Conceptual change. In M. Orey (Ed.), Emerging perspectives on learning, teaching, and technology. http://projects.coe.uga.edu/epltt/

[29] Vosniadou, S. (2002). On the nature of naive physics. In M. Limon & L. Mason (Eds.), Reconsidering conceptual change: Issues in theory and practice (pp. 61-76). Dordrecht, The Netherlands: Kluwer Academic Publishers.

[30] Behar, R. (2001). Aportaciones para la mejora del proceso de enseñanza aprendizaje de la estadística (Doctoral thesis). España: Universidad Politécnica de Cataluña.

[31] Moore, D. (2010). The basic practice of statistics. New York: W.H. Freeman and Company.

[32] Ross, S. (1996). Introductory Statistics. New York: McGraw-Hill.

[33] Christensen, H. (2008). Estadística paso a paso. México: Trillas.

[34] Batanero, C. (2005). Significados de la probabilidad en la educación secundaria. Revista Latinoamericana de Investigación en Matemática Educativa, RELIME, 8(3), 247-263. http://www.redalyc.org/articulo.oa?id=33508302

[35] Fischbein, E. (1987). Intuition in science and mathematics: An educational approach. Dordrecht, The Netherlands: Reidel.

[36] Totohasina, A. (1992). Méthode implicative en analyse de données et application á l'analyse de conceptions d'étudiants sur la notion de probabilité conditionnelle (Doctoral thesis). Université Rennes.