COVARIATIONAL REASONING PATTERNS OF HIGH SCHOOL STUDENTS IN PROBLEMS OF CORRELATION AND LINEAR REGRESSION

PATRONES DE RAZONAMIENTO COVARIACIONAL DE ESTUDIANTES DE BACHILLERATO EN PROBLEMAS DE CORRELACION Y REGRESION LINEAL

Miguel Medina
Centro de Investigación y Estudios Avanzados-IPN
ingecivi.6@gmail.com

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

The topics of correlation and linear regression constitute a complex and subtle system of statistical and mathematical ideas whose teaching-learning raises numerous practical and theoretical problems. In this research paper, the patterns of reasoning that students exhibit, under the approach of informal inferences when they face problems of correlation and regression line are identified. To achieve this, activities were implemented in two stages: in the first stage, two problems (one of estimation and other of best-fit line) were applied to be solved using pencil and paper; the second stage incorporates the use of the Fathom software.

Keywords: Correlation, Linear Regression, Technology, Informal Inferences, Reasoning.

Background and Issues

Covariational statistical reasoning consists of the processes that allow subjects to perceive, describe and justify the relationships between statistical variables. These processes occur in two ways, on the one hand they occur in the subject's mind, and on the other, in the spoken or written discourse that occurs when relationships between variables are described or justified (Moritz, 2004). When the solid arguments or generalizations that students make are highlighted, based on the information they have, instead of only the representations they carry out, we speak of inferences; in this study we will refer to informal statistical inferences (without a formal instruction or explicitly formal procedure). Several authors have reported relatively recently, studies on informal inferential reasoning using the informal statistical inferences made by students as the premise (Ben-Zvi, 2006; Pfannkuch, 2005). For example, in the aforementioned Ben-Zvi work, the informal statistical reasoning of 5th grade students is analyzed and developed within a technological environment, reporting that the use of technological tools showed argumentative advantages in the way students presented ideas.

In this sense, we focus on covariational reasoning starting with informal inferences that students make, taking as reference some relevant clues about this type of reasoning. For example, Zieffler and Garfield (2009, p. 11) summarize some findings: often, students, 1) are significantly influenced by their personal beliefs regarding their covariational judgments; 2) they frequently assume that there is a correlation between two events that is non existent (illusory correlation) ; 3) they are liable to imply causal relationships when dealing with covariation tests; and 4) have difficulty reasoning about covariation when the relationship between variables is negative. Regarding the determination of the line of best fit, some studies, such as those by Casey (2014) and Casey and Wasserman (2015), show that: 1) it is necessary to induce students to convert the data collected in tables into a graphic representation (Dispersion diagram); 2) Students have difficulty observing the global trend of a data set when reading a scatter plot, because they focus their attention on isolated points and perceive the data as a series of individual cases, rather than considering them holistically (with invisible characteristics for isolated points); 3) Many students, when drawing a line of best fit, focus their

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

attention on characteristic points such as the first or last, the highest or the lowest, or a subset of these.

Theoretical framework

This work uses the theory of Rubin, Hammerman and Konold (2006) on informal inferences reasoning seen as statistical reasoning that considers the dimensions of: *Aggregate*, *Signals and Noise*, and *Various Forms of Variability*. An aggregate vision involves the added characteristics of individual cases, that when seen together, enable properties to emerge that are different from those of the individual cases.. Signals and noise refers, on the one hand, to constant elements in statistics such as the mean or the best fit line (signals) and on the other to variable elements that serve to introduce variability around any signal (noise). The idea of considering the various forms of variability refers to the fact that when making judgments from a set of data, the variability that underlies the situation must be considered. Various studiesin the area of statistical education show that students at all levels have difficulties when reasoning about these ideas (Ben-Zvi and Garfield, 2004; Bakker *et al*., 2004).

Based on these ideas, we propose as a central objective of this research: To determine and characterize the reasoning patterns of high school students in the face of correlation problems and the line of best fit, as informal inferences.

Method

First stage of the study

Participants. The questionnaire was applied to a group of 30 high school students (16 to 18 years old) who were taking the subject of Statistics and Probability at the College of Sciences and Humanities, Plantel Vallejo, located in Mexico City. Two instructor-researchers, co-authors of this article participated in the application of the questionnaire; they provided worksheets containing the problems and guided the dynamics of the sessions, in particular clarifying doubts about the instructions of the activities without delving into the content.

Instruments and Execution. The questionnaire was applied to students who had not received formal and expository instruction on these topics. Two problems were chosen to report on, which are shown in Figure 1. The first priblem is an Estimation problem taken from Moore (1988) and which was modified to the present version. The second is a best fit line problem for which a scatter diagram of the variables Height and Weight is presented, in a situation in which they are directly correlated and with a linear trend. Given the diagnostic nature of the questionnaire, a didactic sequence was not elaborated as such, instead the students were given a series of problems that were solved individually during two sessions of 90 min each, only allowing the use of pencil and paper.

Estimation Problem

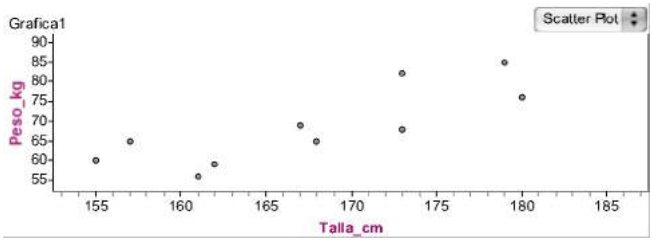
The Morales family is about to install solar panels at their home to cut down on heating costs. To better understand the savings that installing these panels can mean, the Morales have been recording their gas consumption for the last year and a half. The table shows the data with the average gas consumption and the average ambient temperature for each month:

Average Temperature (° C)	5.2	-9.8	-5.4	0.2	4.1	11.3
Gas Consumption (m^3)	17.6	30.5	24.9	21	14.8	11.2
Average Temperature (° C)	16.3	18.5	18.5	18	15.2	11.8
Gas Consumption (m^3)	4.8	3.4	3.4	3.4	5.9	8.7
Average Temperature (° C)	1.8	0.7	-10.4	1.8	17.1	10.7
Gas Consumption (m^3)	17.9	20.2	30.8	19.3	6.3	10.7

If the average temperature recorded in a month is 8 °C, what is the gas consumption expected by the Morales family in that month? Explain your answer:

Line of Best Fit Problem

The graph below shows the height and weight data for 10 high school students. Draw the line that you think best fits the data.



Explain the criteria you used to draw the line:

Figure1. Estimation and Best Fit Line Problem

Analysis Methodology. The analysis of the responses was carried out through a coding process (types of reasoning or inferences) of the responses, gradually refining through comparative analysis. Each researcher carried out an open coding of the data, generating codes that represent patterns of reasoning according to the similarities in the procedures and arguments declared by the students, which were then compared with each other. Comparison of these

coding proposals resulted in a more consistent and abstract set of codes than the descriptions initially developed; this preliminary scheme facilitated inducing some central ideas of the students' covariational reasoning.

Second stage of the investigation

Participants. The activities were applied to a group of 40 high school students (20 couples) (16 to 18 years old) from the Colegio de Bachilleres Plantel 2, which is also located in Mexico City. Given the uncontrollable nature of the classroom, certain students were absent from some class sessions. Similarly, the two researchers involved in the first stage of the study participated in the experiment. In addition to guiding the class sessions, their participation included supporting the students in manipulating the Fathom software. As in the previous stage, the professor-researchers were prevented from delving into the thematic content.

Instruments and Execution. Again, for this report we have chosen two problems that we want to focus on: the first, which deals with the estimation of a response value, which was the same as in the first stage, only in this case the use of the Fathom software was incorporated as a tool for the student to be able to build a scatter diagram on and obtain the least squares line, to make the requested estimate. The second (figure 2) also deals with proposing the line that best fits the data, but the context of the situation was modified. Unlike the first stage, in this stage the students collected statistical data by measuring some of their own physiological attributes (namely, for the same student, the measurement of their height against that of their arm). In addition, they used the software to draw and manipulate a line that they considered best fit the data.

Line of Best Fit Problem

According to studies of the anatomy of the human body, there is a certain relationship between the height of people and the measurement of some parts of the body.

With the data of the measurement of the arms (from elbow to shoulder) and the height of your colleagues, which were collected in the first work session, and which are shown in the table, answer what is asked of you.

X (arm, cm)	30	33	35	36	32	38
Y (height, cm)	153	164	175	177	160	175
X (arm, cm)	34	31	35	29	28	38
Y (height, cm)	167	154	180	162	155	180

Figure 2. Problem of best fit line.

Analysis Methodology. The coding process of the data obtained in this stage was extended to the search for solid and consistent relations between the codes identified in both phases of the study; this process was carried out by investigating possible coincidences and patterns that we considered emerged with sufficient sense and coherence. Once the responses from the two stages of the study were coded, conceptual connections were identified between the proposed codes, to finally define the reasoning patterns as informal inferences made by the students.

Results

Analysis of the first stage of the study

During this stage, a system of codes representing the types of reasoning or inferences showing the student responses was developed, these codes are defined in terms of the arguments exhibited, that is to say the identification correlation between variables, use of all available data, or perception of the uncertainty that underlies the data. Some evidence of student responses is presented.

For the problem of estimation, a type of reasoning was defined as *Arithmetic Interpolation* (4/28 responses) representing the responses where the student takes a range of values in which the data of the given temperature (8 ° C) is included and with its corresponding gas consumptions, the student obtains the average to make the estimate. In this case the student uses data from the two variables, which we consider a slight statistical approach towards obtaining the arithmetic average. The response of the student E10 is presented as evidence in figure 3, where it can be seen that he chooses two temperature values (5.2 ° C and 10.7 ° C) among which is the data of 8 ° C - of which the estimate is requested - and with their corresponding gas consumption values (17.6 m^3 and 10.7m^3) calculate their average to give their estimated answer.

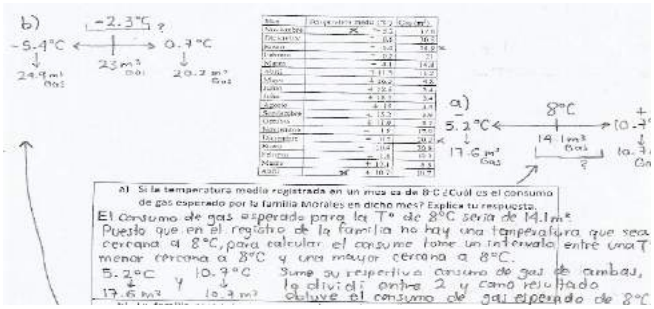


Figure 3. Student response E10. Arithmetic Code Interpolation

The following code we call *Proportional Arithmetic* (8/28) includes the answers where students look for a proportionality factor; choosing a pair of data (X-Temperature, Y-Consumption) and with the given temperature of 8 ° they form a rule of three, assuming that there is a proportional relationship between the variables. The student's answer E9 is shown as an example, in which he chooses the pair of values that correspond to the first month of the table (5.2 ° C, 17.6m^3) and with the value of 8 ° C he forms a rule of three.

Figure 4. Student answer E9. Proportional Arithmetic Code

The *Arithmetic Reasoning Following A Pattern* (2/28) includes the answers where the students take as a reference the given temperature value (8 °) and try, from the data in the table, to "complete" this value by means of some arithmetic procedure, and once they get it they use that data in order to obtain their answer.

In the *Arithmetic Type Without Defined Pattern* (6/28) the students used some basic operations (addition, subtraction, multiplication and division; in one case the square root is used) but without being able to deduce a well-defined procedure.

In the *Perception Of The Trend* code (6/28) the student does not carry out any operation and only focuses his attention on the data in the table and his answers are based on a visual analysis of the trend of the data, in particular in the meaning of their behavior.

Finally, the code *Without Argument* (2/28) was defined, which represents those responses where the student only provides the result, without arguing or making their procedure explicit.

For the problem of drawing the line of best fit, two characteristic types of reasoning were found, the *Partition* code (6/21 responses) where the responses were classified in which the student draws the line or refers to the fact that their position must be such that passes through the middle of the cloud of points, following its direction, that is to say, traced diagonally, leaving the same number of points on one side and the other of the line, as shown in Figure 5.

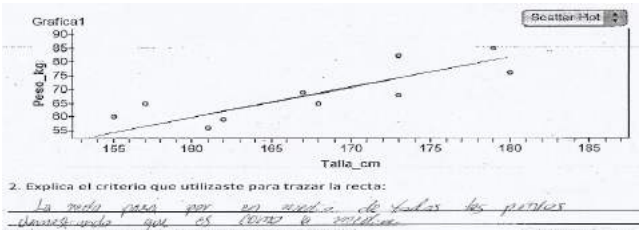


Figure 5. Student response E14. Partition Code

The *Belonging* code (15/21) includes the responses that show two types of behavior, on the one hand, those where the student draws or refers to the fact that the line must pass through as many points as possible or through all of them, and there are also the answers where the student draws the line insuring that it passes through specific points of the cloud, as in the answer shown in the following Figure, where it is argued that it must pass through two points (the lowest and the highest).

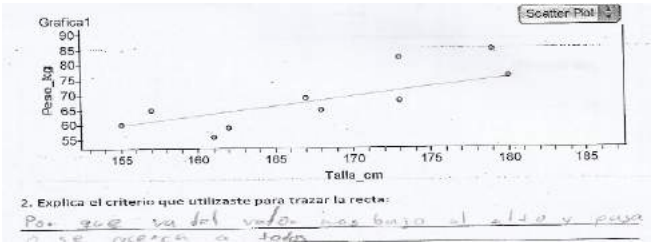


Figure 6. Student response E28. Membership Code

Analysis of the second stage of investigation

For the estimation problem, most of the identified reasoning was presented in the same way as in the diagnostic questionnaire, however, at this stage it stands out that the Arithmetic Interpolation code is absent in the students' responses and was replaced by the Use code of Software (3/15 pairs), in which the student uses the software to modify the position of a point in the cloud up to the given temperature value (8 ° C) and, following the trend of the data set, provides the estimated value of gas consumption. An evidence of this code is shown below:

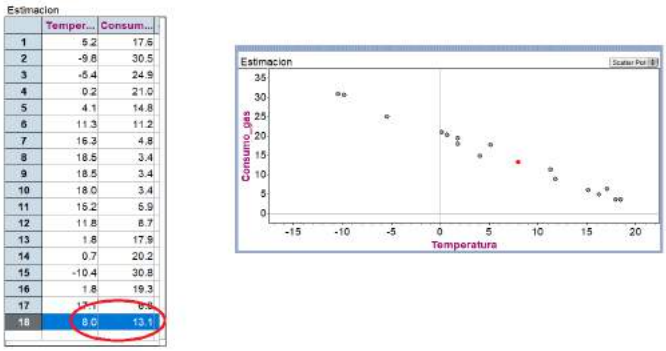


Figure 7. P4 partner response. Software Use Code

In the problem of the line of best fit, some students took the measurement of the arm (from elbow to shoulder) and with its corresponding height they formed a new bivariate database; It is from this set that the respective scatter diagram with which the students worked with was constructed.

The *Partition* and *Belonging* codes emerged with a frequency similar to that observed in the previous stage, with the exception of a new argument that we called *Closeness* (7/19 couples); The code includes the answers in which the students position the line with the help of the software in such a way that it is as close as possible to most of the points. The response of the pair of students P6 is included as evidence.

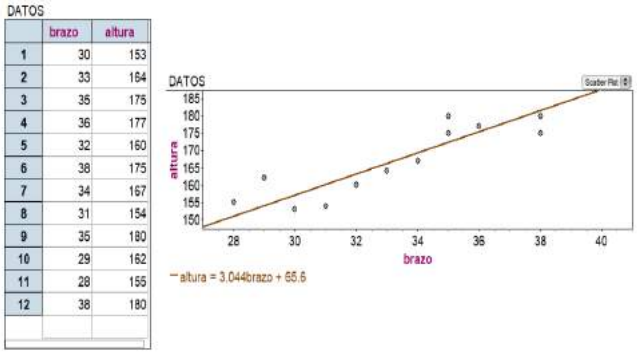


Figure 8. Pair answer P6. Closeness Code

A summary table of the analysis and initial coding is shown for the two stages of the study:

Table 1. Coding Process

Code	First stage	Second stage	Relative frequency	
Arithmetic Interpolation	✓		0.143	0.000
Use of Software		✓	0.000	0.200
Arithmetic Proportionality	✓	✓	0.286	0.267
Arithmetic Following a Pattern	✓	✓	0.071	0.267
Arithmetic Without Following a Pattern	✓	✓	0.214	0.267
Perception of the Trend	✓		0.214	0.000
No Argument	✓		0.071	0.000
Partition	✓	✓	0.286	0.211
Closeness		✓	0.000	0.368
Belonging	✓	✓	0.714	0.421

The next phase of the analysis was to compare the codes defined in the two stages of the study and to identify common covariational reasoning traits and some arguments raised in the second

stage; the process of conceptualization of reasoning (inference) of student responses are described below.

Based on the inferences made by the students, conceptual connections were identified between the reasoning they exhibit, hence the codes *Perception of Tendency* and *Use of Software* (estimation problem) were reclassified in the reasoning pattern *Notion of Aggregate,* since in both students use the complete set of values that they are given to argue their responses; regardless of whether or not their results are normatively correct. On the other hand, the codes *Interpolation*, *Proportionality*, *Arithmetic following a pattern, Arithmetic without following a Patern* (estimation problem), *Partition* and *Belonging* (problem straight adjustment) will be reclassified as a *search for a signal*. The interconnection between these codes lies in the fact that the students infer that there must be a kind of clue to solve the problem, and they carry out a search in the data that was provided. It seems that the student suspects that there is a hidden pattern or structure in a subgroup of data (he does not use the totality of the data or contemplate the set of points) that will lead him to constant structures, absent of uncertainty or variation, that are familiar to him. Finally, the *Closeness* code (best fit line problem) was renamed as *Sense of Variability*, since features are perceived, albeit in a spurious way, that students infer a relationship between the fit model and the points of the cloud, in the scatter diagram. The *No Argument* code that was presented in the estimation problem, by not providing evidence of some type of inference made by the students, was not considered at this stage of the analysis.

Table 2. Process Conceptualization

Code	Informal Reasoning Pattern / Inferences
Perception of the Trend Use of Software	Notion of Aggregate
Arithmetic Interpolation Arithmetic Proportionality Arithmetic Following a Pattern Arithmetic Without Following a Pattern Partition Belonging	Search for a Signal
Closeness	Sense of Variability

Table 2 shows a summary of the conceptualization and definition of the identified reasoning patterns.

Conclusions

From our theoretical perspective, an informal inference is a type of reasoning that includes considerations in several dimensions (Rubin, Hammerman and Konold, 2006) and in specific study conditions. In our case a first dimension is the *Notion Added*, where a holistic perception of the problem situation over the contemplation of individual cases is privileged. In this sense, the initially defined codes, *Perception of the trend* and *Use of software*, present as a pattern that the students consider all the data they have available.On the one hand, they estimate the value of the requested response variable (gas consumption) by making a purely visual analysis based on the trend of the data set, that is, they argue their response based on the global behavior of the set of values. On the other hand, when they use Fathom, when constructing the scatter diagram and observing the trend of the point cloud again, they identify that there must be a certain relationship between the points, so they choose one and modify its position in such a way that the value matches the independent variable (8 ° C temperature) with that of the corresponding response variable (gas consumption), following the general "shape" of the cloud. Although it is true that in none of the above cases the

response of the students is normatively adequate, it does allow us to appreciate that under certain circumstances they are able to perceive that in bivariate relationships it is necessary to consider the characteristics and behavior of the data as a whole and, based on this, infer the value of some particular point of interest.

Another dimension to consider is reflected with the pattern that we define as *Search for a signal* , where the answers show that given the difficulty represented by the uncertainty or the intrinsic variation in this type of problem, the students try to solve the situation in a familiar terrain for them or with which they feel comfortable, possibly for this reason they mostly use arithmetic procedures (rule of three, proportionality factor or additions and subtraction) to make the estimation; where they also only use part of the data. It is also the case of the best fit line problem, in which in the absence of the notion of uncertainty or of this aggregate view, they partially use the available data, referring to the linear function model as an alternative to fit a line to a distribution of points that presents a linear trend, considering only some of these points in their choice, ignoring the influence of all of these and their joint variation, and above all, defining two types of data: those that do or do not belong to the linear model that they choose to plot.

As a third dimension we propose, *Sense of Variability*, represented by the code that was initially defined as *Closeness*. In this pattern of reasoning, the answers that refer, albeit briefly, to the perception that there is some variation that underlies these types of problems and that that must be considered when proposing a line of adjustment for a set of points., just as the students did when plotting and arguing that the line should be positioned as close to most of the points as possible.

We trust that the identification of these reasoning patterns as the way in which students make informal inferences in the face of statistical association problem situations adds to the body of knowledge in the study of bivariate data, without neglecting the importance of exploring obstacles of learning, such as the apparent disconnection that the student has between the predictive or inferential nature inherent in the linear regression model and its identification as the line that best fits the set of points, as well as the difficulty to conceive the data set as an aggregate , that is, as a system in which they are linked to each other and have the property of being deviations from the same model.

References

Bakker, A., Biehler, R., and Konold, C. (2004). Should Young Students Learn About Box Plots? In G. Burrill and M. Camden (Eds.), *Proceedings of IASE 2004 Roundtable on Curricular Development in Statistics Education*, Lund, Sweden. Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D., and Garfield, JB (Eds.) (2004). *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* . Dordrecht, the Netherlands: Kluwer Academic Publishers.

Ben-Zvi, D. (2006). Scaffolding student's informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education*: Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Brazil. Downloaded from http://www.stat.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf

Casey, S. (2014). Teachers 'knowledge of students' conceptions and their development when learning linear regression, in K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics. ICOTS9 Invited Paper-Refereed*. Voorburg, The Netherlands: International Statistical Institute

Casey, S., & Wasserman, N. (2015), "Teachers' Knowledge about Informal Line of Best Fit." *Statistics Education Research Journal*, *14* (1), 8-30.

Moore, DS (1988). Should mathematicians teach statistics? *The College Mathematics Journal*, *19* (1), 3–7.

Moritz, JB (2004). Reasoning about covariation. In D. Ben-Zvi, & J. Garfield (Eds.) *The challenge of developing statistical literacy, reasoning and thinking* (pp. 227-256). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pfannkuch, M. (2005). Informal inferential reasoning: A case study. In K Makar (Ed.), Proceedings of the International Forum for Research in Statistical Reasoning, Thinking and Literacy . Auckland, NZ. Brisbane: University of Queensland.

Rubin, A., Hammerman, J., and Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman and B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg: The Netherlands: International Statistical Institute.

Zieffler, AS, & Garfield, J. (2009). Modeling the Growth of Students' Covariational Reasoning during an introductory statistics course. *Statistics Education Research Journal, 8* (1), 7-3.

PATRONES DE RAZONAMIENTO COVARIACIONAL DE ESTUDIANTES DE BACHILLERATO EN PROBLEMAS DE CORRELACION Y REGRESION LINEAL

COVARIATIONAL REASONING PATTERNS OF HIGH SCHOOL STUDENTS IN PROBLEMS CORRELATION AND LINEAR REGRESSION

Miguel Medina
Centro de Investigación y Estudios Avanzados-IPN
ingecivi.6@gmail.com

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

Los temas de correlación y regresión lineal constituyen un sistema complejo y sutil de ideas estadísticas y matemáticas cuya enseñanza-aprendizaje plantea numerosos problemas prácticos y teóricos. En esta investigación se identifican patrones de razonamiento que exhiben los estudiantes, bajo el enfoque de inferencias informales cuando enfrentan problemas de correlación y regresión lineal; para esto se implementaron actividades en dos etapas: en la primera se aplicaron dos problemas (uno de estimación y otro de recta de mejor ajuste) para ser resueltos a lápiz y papel; en la segunda se incorpora el uso del software Fathom.

Palabras clave: Correlación, Regresión Lineal, Tecnología, Inferencias Informales, Razonamiento.

Antecedentes y Problemática

El razonamiento estadístico covariacional consiste en los procesos que le permiten a los sujetos percibir, describir y justificar las relaciones entre variables estadísticas; estos procesos se presentan en dos sentidos, por un lado ocurren en la mente del sujeto, y por otro, en el discurso hablado o escrito cuando se describen o justifican relaciones entre variables (Moritz, 2004). Cuando se destacan las argumentaciones sólidas o generalizaciones que los estudiantes realizan, a partir de la información con que cuentan, en lugar de solo las representaciones que llevan a cabo se habla de inferencias: en este estudio nos referiremos a las inferencias estadísticas informales (sin una instrucción formal o procedimiento explícitamente formales). Varios autores han reportado en fechas relativamente recientes, estudios sobre el razonamiento inferencial informal tomando como premisa las inferencias estadísticas informales que realizan los estudiantes (Ben-Zvi, 2006; Pfannkuch, 2005). Por ejemplo en el trabajo de Ben-Zvi mencionado, se analiza y desarrolla el razonamiento estadístico informal de estudiantes de 5° dentro de un ambiente con tecnología, reportando que el uso de herramientas tecnológicas mostró ventajas argumentativas en la presentación de ideas por parte de los estudiantes.

En este sentido, nos enfocamos en el razonamiento covariacional partiendo de inferencias informales que los alumnos hacen, tomando como referencia algunas pistas relevantes sobre este tipo

de razonamiento; por ejemplo, Zieffler y Garfield (2009, p. 11) resumen algunos hallazgos: a menudo, los estudiantes, 1) se encuentran influenciados significativamente por sus creencias personales con respecto a sus juicios covariacionales; 2) suponen frecuentemente que existe correlación entre dos eventos que no lo están (correlación ilusoria); 3) son susceptibles a implicar relaciones causales cuando tratan con pruebas de covariación; y 4) tienen dificultad para razonar acerca de la covariación cuando la relación entre las variables es negativa. Sobre la determinación de la recta de mejor ajuste, algunos estudios, como los de Casey (2014) y Casey y Wasserman (2015), manifiestan que: 1) es necesario inducir a los estudiantes convertir los datos recopilados en tablas en una representación gráfica (diagrama de dispersión); 2) los estudiantes tienen dificultad para observar la tendencia global de un conjunto de datos cuando leen un diagrama de dispersión, debido a que enfocan su atención en puntos aislados y perciben los datos como una serie de casos individuales, en lugar de considerarlos de manera holística (con características invisibles para puntos aislados); 3) muchos estudiantes, al trazar una recta de ajuste, enfocan su atención en puntos característicos como el primero o el último, el más alto o más bajo, o en un subconjunto de estos.

Marco Teórico

Este trabajo toma ideas de la teoría de Rubin, Hammerman y Konold (2006) sobre inferencias informales vistas como razonamientos estadísticos que implican considerar las dimensiones de: *Agregado*, *Señales y Ruido*, y *Diversas Formas de Variabilidad*. Una visión de agregado implica que las características de los casos individuales, al ser vistas en conjunto, permiten que emerjan propiedades que son diferentes de las que tienen los casos individuales por sí mismos. Señales y ruido, se refiere por un lado a elementos constantes en estadística como la media o la recta de ajuste (señales) y por otro a elementos variables que sirven para introducir variabilidad alrededor de cualquier señal (ruido). La idea de considerar las diversas formas de variabilidad, se refiere a que al elaborar juicios a partir de un conjunto de datos se debe considerar la variabilidad que subyace en la situación. Diversas investigaciones en el área de la educación estadística manifiestan que estudiantes de todos los niveles presentan dificultades al razonar sobre estas ideas (Ben-Zvi y Garfield, 2004; Bakker *et al*., 2004).

Con base en estas ideas planteamos como objetivo central de la investigación: Determinar y caracterizar los patrones de razonamiento de estudiantes de bachillerato ante problemas de correlación y de recta de mejor ajuste, como inferencias informales.

Método

Primera etapa de la investigación

Participantes. El cuestionario se aplicó a un grupo de 30 alumnos de bachillerato (16 a 18 años de edad) que se encontraban cursando la asignatura de Estadística y Probabilidad y pertenecientes al Colegio de Ciencias y Humanidades, Plantel Vallejo, localizado en la Ciudad de México. Participaron en la aplicación del cuestionario dos profesores-investigadores coautores del presente artículo; proporcionaron hojas de trabajo que contenían los problemas y guiaron la dinámica de las sesiones, en particular aclarando dudas sobre la instrucción de las actividades sin profundizar en los contenidos.

Instrumentos y Ejecución. El cuestionario se aplicó a estudiantes que no habían recibido instrucción formal y expositiva sobre dichos temas. Se eligieron dos problemas para reportar que se muestran en la figura1, el primero es un problema de Estimación tomado de Moore (1988) y que se modificó a la presente versión. En el segundo problema, de recta de ajuste, se presenta un diagrama de dispersión de las variables Talla y Peso, en una situación en que están correlacionadas directamente y con una tendencia lineal. Dado el carácter diagnóstico del cuestionario, no se elaboró como tal una secuencia

didáctica, sino una serie de problemas que fueron resueltos individualmente por los alumnos durante dos sesiones de 90 min cada una, permitiéndose únicamente el uso de lápiz y papel.

Problema de Estimación

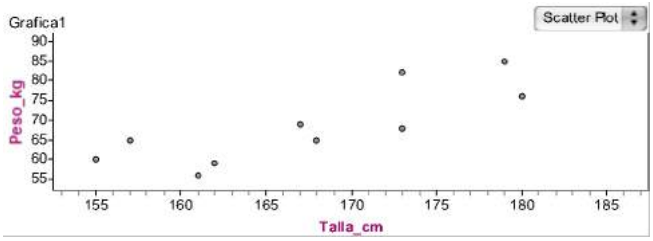
La familia Morales está a punto de instalar paneles solares en su casa para reducir el gasto en la calefacción. Para conocer mejor el ahorro que puede significar la instalación de dichos paneles, los Morales han ido registrando su consumo de gas durante el último año y medio. En la tabla se muestran los datos con el promedio del consumo de gas y de la temperatura media ambiental de cada mes:

Temperatura	5.2	-9.8	-5.4	0.2	4.1	11.3	16.3	18.5	18.5
Consumo de Gas (m^3)	17.6	30.5	24.9	21	14.8	11.2	4.8	3.4	3.4
Temperatura	18	15.2	11.8	1.8	0.7	-	1.8	17.1	10.7
Consumo de Gas (m^3)	3.4	5.9	8.7	17.9	20.2	30.8	19.3	6.3	10.7

Si la temperatura media registrada en un mes es de 8°C ¿Cuál es el consumo de gas esperado por la familia Morales en dicho mes? Explica tu respuesta:

Problema de Recta de Mejor Ajuste

La siguiente gráfica muestra los datos de la talla y el peso de 10 estudiantes de bachillerato. Traza la recta que piensas se ajusta mejor a los datos.



Explica el criterio que utilizaste para trazar la recta:

Figura 1. Problema de Estimación y de Recta de mejor ajuste

Metodología de Análisis. El análisis de las respuestas fue llevado a cabo mediante un proceso de codificación (tipos de razonamiento o inferencias) de las respuestas, refinándose gradualmente a través de un análisis comparativo. Cada investigador realizó una codificación abierta de los datos generando códigos que representan patrones de razonamiento según las semejanzas en los procedimientos y argumentos declarados por los estudiantes, que luego fueron comparados entre sí. La comparación de estas propuestas de codificación dio lugar a un conjunto de códigos más consistente y abstracto que las descripciones inicialmente elaboradas; este esquema preliminar facilitó inducir algunas ideas centrales en el razonamiento covariacional de los estudiantes.

Segunda etapa de la investigación

Participantes. Las actividades se aplicaron a un grupo de 40 estudiantes (20 parejas) de bachillerato (16 a 18 años) del Colegio de Bachilleres Plantel 2, que también se encuentra ubicado en la Ciudad

de México; dada la naturaleza poco controlable del aula escolar, ciertos estudiantes se ausentaron de algunas sesiones de clase. De igual manera, en el experimento participaron los dos investigadores involucrados en la primera etapa del estudio; además de guiar las sesiones de clase, su participación incluyó el apoyar a los alumnos en la manipulación del software Fathom. Al igual que en la etapa previa, se evitó que los profesores-investigadores profundizaran en los contenidos temáticos.

Instrumentos y Ejecución. Nuevamente, para este reporte hemos elegido dos problemas en los que deseamos enfocarnos: el primero, que trata sobre la estimación de un valor respuesta, fue el mismo que en la primera etapa, solo que en este caso se incorporó el uso del software Fathom como una herramienta para que el estudiante estuviera en posibilidad de construir un diagrama de dispersión y obtener la recta de mínimos cuadrados, para realizar la estimación pedida. El segundo (figura2) también trata sobre proponer la recta que mejor ajusta a los datos pero el contexto de la situación fue modificado. A diferencia de la primera etapa, en esta los estudiantes recolectaron los datos estadísticos a través de la medición de algunos atributos fisiológicos propios (a saber, para un mismo estudiante, la medida de su talla contra la de su brazo); además, utilizaron el software para trazar y manipular una recta que consideraran se ajustara mejor a los datos.

Problema de Recta de Mejor Ajuste

Según estudios de la anatomía del cuerpo humano, existe cierta relación entre la talla (altura) de las personas y la medida de algunas partes del cuerpo.

Con los datos de la medida de los brazos (del codo al hombro) y de la altura de tus compañeros, que se recolectaron en la primera sesión de trabajo, y que se muestran en la tabla, contesta lo que se te pide.

X (brazo,cm)	30	33	35	36	32	38	34	31	35	29	28	38
Y (altura,cm)	153	164	175	177	160	175	167	154	180	162	155	180

Figura 2. Problema de Recta de Mejor Ajuste

Metodología de Análisis. El proceso de codificación de los datos obtenidos en esta etapa se extendió a la búsqueda de relaciones sólidas y consistentes entre los códigos identificados en ambas etapas del estudio; dicho proceso se realizó indagando posibles coincidencias y patrones que consideramos emergieron con suficiente sentido y coherencia. Una vez que se codificaron las respuestas de las dos etapas del estudio, se identificaron conexiones conceptuales entre los códigos propuestos, para finalmente definir los patrones de razonamiento como inferencias informales que realizan los alumnos.

Resultados

Análisis de la primera etapa de investigación

Durante esta etapa se desarrolló un sistema de Códigos que representan los tipos de razonamientos o inferencias que muestran las respuestas de los estudiantes, estos códigos se encuentran definidos en función de las argumentaciones que exhiben, es decir, la identificación de la correlación entre las variables, la utilización de todos los datos disponibles o la percepción de la incertidumbre que subyace en los datos. Se presentan algunas evidencias de las respuestas de los estudiantes.

Para el problema de estimación, un tipo de razonamiento se definió como *Aritmético Interpolación* (4/28 respuestas) que representa las respuestas donde el estudiante toma un intervalo de valores en el que se encuentra incluido el dato de la temperatura dado (8°C) y con sus correspondientes consumos de gas obtiene el promedio para realizar la estimación. En este caso el alumno utiliza datos de las dos variables, lo que consideramos un ligero acercamiento estadístico al obtener el promedio aritmético.

Se presenta como evidencia la respuesta del estudiante E10 en la figura3, donde se aprecia que elige dos valores de temperatura (5.2°C y 10.7°C) entre los que se encuentra el dato de 8°C –del que se pide la estimación- y con sus correspondientes valores de consumo de gas (17.6 m^3 y 10.7m^3) calcula su promedio para dar su respuesta estimada.

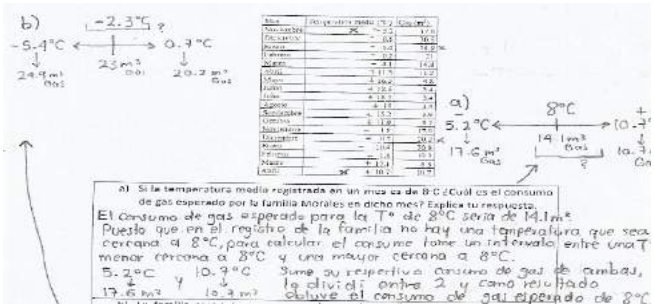


Figura 3. Respuesta estudiante E10. Código Aritmético Interpolación

El siguiente código lo llamamos *Aritmético Proporcional* (8/28) incluye las respuestas donde los estudiantes buscan un factor de proporcionalidad; eligiendo una pareja de datos (X-Temperatura, Y-Consumo) y con la temperatura dada de 8° forman una regla de tres, asumiendo que entre las variables existe una relación proporcional. Se muestra como ejemplo la respuesta del estudiante E9, en la que elige la pareja de valores que corresponden al primer mes de la tabla (5.2°C, 17.6m^3) y con el valor de 8°C forma una regla de tres.

Figura 4. Respuesta estudiante E9. Código Aritmético Proporcional

El razonamiento *Aritmético Siguiendo Un Patrón* (2/28) incluye las respuestas donde los estudiantes toman como referencia el valor de la temperatura dado (8°) y tratan, a partir de los datos de la tabla, de "completar" este valor mediante algún procedimiento aritmético, y una vez que lo consiguen utilizan esos datos para obtener su respuesta.

En el tipo *Aritmético Sin Patrón Definido* (6/28) los estudiantes utilizaron algunas operaciones básicas (suma, resta, multiplicación y división; en un caso se utiliza la raíz cuadrada) pero sin que se pueda deducir un procedimiento bien definido.

En el código *Percepción De La Tendencia* (6/28) el estudiante no realiza operación alguna y sólo enfoca su atención en los datos de la tabla y sus respuestas se basan en un análisis visual de la tendencia de los datos, en particular en el sentido de su comportamiento.

Finalmente, se definió el código *Sin Argumento* (2/28) que representa aquellas respuestas donde el estudiante solo aporta el resultado, sin argumentar o hacer explícito su procedimiento.

Para el problema de trazar la recta de mejor ajuste se encontraron dos tipos de razonamiento característicos, el código *Partición* (6/21 respuestas) donde se clasificaron las respuestas en las que el estudiante traza la recta o hace referencia a que su posición debe ser tal que pase por en medio de la nube de puntos, siguiendo la dirección de ésta, es decir trazada de forma diagonal, dejando de un lado y del otro de la recta, el mismo número de puntos, como lo muestra la Figura 4.

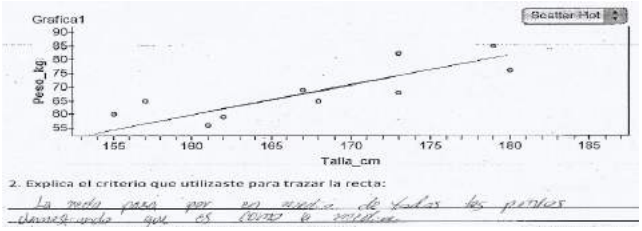


Figura 5. Respuesta estudiante E14. Código Partición

En el código *Pertenencia* (15/21) se incluyen las respuestas que muestran dos tipos de comportamiento, por un lado aquellas donde el estudiante traza o refiere que la recta debe pasar por el mayor número posible de puntos o por la totalidad de estos, y también están las respuestas donde el estudiante traza la recta cuidando que pase a través puntos específicos de la nube, como en la respuesta mostrada en la siguiente Figura, donde se argumenta que debe pasar por dos puntos (el más bajo y más alto).

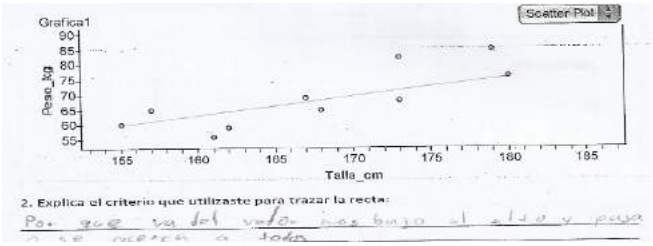


Figura 6. Respuesta estudiante E28. Código Pertenencia

Análisis de la segunda etapa de investigación

Para el problema de estimación, la mayoría de los razonamientos identificados se presentaron de igual forma que en el cuestionario diagnóstico, sin embargo en esta etapa destaca que el código *Aritmético Interpolación* se encuentra ausente en las respuestas de los estudiantes y fue sustituido por el código *Uso de Software* (3/15 parejas), en el cual el estudiante utiliza el software para modificar la posición de un punto de la nube hasta el valor de temperatura dada (8° C) y, siguiendo la tendencia del conjunto de datos, proporciona el valor estimado del consumo de gas. Una evidencia de este código se muestra a continuación:

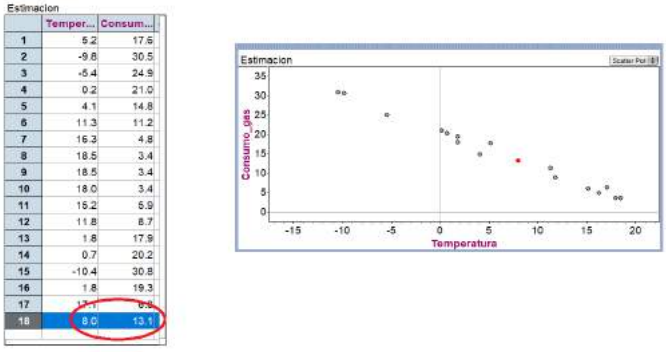


Figura 7. Respuesta pareja P4. Código Uso de software

En el problema de recta de mejor ajuste algunos alumnos se tomaron la medida del brazo (del codo al hombro) y con su correspondiente altura conformaron una nueva base de datos bivariados; a partir

de este conjunto es que se construyó el respectivo diagrama de dispersión con el que trabajaron los estudiantes.

Los códigos *Partición* y *Pertenencia* emergieron con una frecuencia similar a la observada en la etapa previa, con la excepción de un nuevo argumento que denominamos como *Cercanía* (7/19 parejas); el código engloba a las respuestas en las que los estudiantes posicionan la recta con ayuda del software de tal manera que se encuentre lo más cerca posible de la mayoría de los puntos. Se incluye como evidencia la respuesta de la pareja de estudiantes P6.

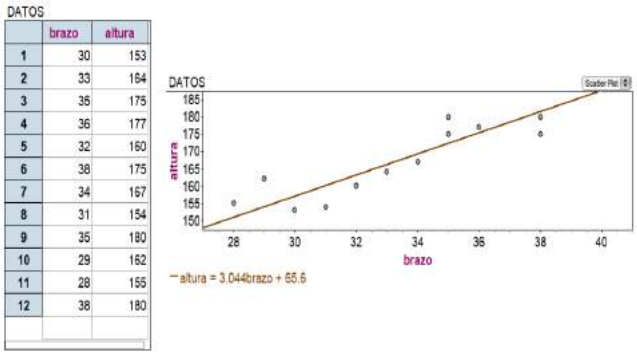


Figura 8. Respuesta pareja P6. Código Cercanía

Se muestra una tabla resumen del análisis y codificación inicial, para las dos etapas del estudio:

Tabla1. Proceso de Codificación

Código	Primera Etapa	Segunda Etapa	Frecuencia Relativa	
Aritmético Interpolación	✓		0.143	0.000
Uso de Software		✓	0.000	0.200
Aritmético Proporcionalidad	✓	✓	0.286	0.267
Aritmético Siguiendo un Patrón	✓	✓	0.071	0.267
Aritmético Sin Seguir un Patrón	✓	✓	0.214	0.267
Percepción de la Tendencia	✓		0.214	0.000
Sin Argumento	✓		0.071	0.000
Partición	✓	✓	0.286	0.211
Cercanía		✓	0.000	0.368
Pertenencia	✓	✓	0.714	0.421

La siguiente fase del análisis consistió en comparar los códigos definidos en las dos etapas del estudio e identificar rasgos de razonamiento covariacional comunes, así como algunos razonamientos que surgieron en la segunda etapa; el proceso de conceptualización de los razonamientos (inferencias) de las respuestas de los estudiantes a continuación se describe.

En función de las inferencias que realizan los estudiantes se identificaron conexiones conceptuales entre los razonamientos que exhiben, de ahí que los códigos *Percepción de la Tendencia* y *Uso de* Software (problema de estimación) se reclasificaron en el patrón de razonamiento Noción de *Agregado* pues en ambos casos los estudiantes hacen referencia a la utilización del conjunto completo de valores de los que disponen, para argumentar sus respuestas; independientemente de que sus resultados sean o no correctas normativamente. Por otro lado los códigos *Interpolación*, *Proporcionalidad*, *Aritmético Siguiendo un Patrón, Aritmético Sin Seguir un Patrón* (problema de estimación), *Partición* y *Pertenencia* (problema de recta de ajuste) se reclasificaron como *Búsqueda de una Señal*. La interconexión entre estos códigos radica en que los estudiantes infieren que debe existir una especie de clave o pista para resolver el problema, y realizan una búsqueda en los datos que se le proporcionaron. Parece que el alumno sospecha que existe un patrón o estructura ocultos en

un subgrupo de datos (no utiliza la totalidad de datos ni contempla el conjunto de puntos) que lo llevarán a estructuras constantes, ausentes de incertidumbre o variación, que le son familiares. Finalmente se renombró al código *Cercanía* (problema de recta de ajuste) como *Sentido de Variabilidad,* ya que se perciben rasgos, aunque de manera espuria, de que los estudiantes infieren una relación presente entre el modelo de ajuste y los puntos de la nube, en el diagrama de dispersión. El código *Sin Argumento* que se presentó en el problema de estimación, al no aportar evidencias de algún tipo de inferencia realizada por los alumnos, no se consideró en esta etapa del análisis.

Tabla 2. Proceso de Conceptualización

Código	Patrón de Razonamiento /Inferencias Informales
Percepción de la Tendencia Uso de Software	Noción de Agregado
Aritmético Interpolación Aritmético Proporcionalidad Aritmético Siguiendo un Patrón Aritmético Sin Seguir un Patrón Partición Pertenencia	Búsqueda de una Señal
Cercanía	Sentido de Variabilidad

La tabla 2 muestra un resumen de la conceptualización y definición de los patrones de razonamiento identificados.

Conclusiones

Desde la perspectiva teórica que abordamos, una inferencia informal es un tipo de razonamiento que incluye consideraciones en varias dimensiones (Rubin, Hammerman y Konold, 2006) y en condiciones específicas de estudio. En nuestro caso una primera dimensión es la *Noción de Agregado*, donde se privilegia una percepción holística de las situación-problema por sobre la contemplación de los casos individuales. En este sentido los códigos inicialmente definidos, Percepción de la tendencia y Uso de software, presentan como patrón el que los estudiantes consideran la totalidad de los datos con que disponen; por un lado hacen la estimación del valor de la variable respuesta pedido (consumo de gas) haciendo un análisis puramente visual a partir de la tendencia del conjunto de datos, es decir que argumentan su respuesta con base en el comportamiento global del conjunto de valores. Por otro lado cuando utilizan Fathom, al construir el diagrama de dispersión y observar nuevamente la tendencia de la nube de puntos, identifican que debe existir cierta relación entre los puntos, por lo que eligen uno y modifican su posición de tal manera que coincida el valor dado de la variable independiente (8°C de temperatura) con el de la variable respuesta (consumo de gas) correspondiente, siguiendo la "forma" general de la nube. Si bien es cierto que en ninguno de los casos anteriores la respuesta de los estudiantes es la normativamente adecuada, si permite apreciar que bajo ciertas circunstancias son capaces de percibir que en las relaciones bivariadas es necesario considerar las características y comportamiento de los datos como un todo y, con base en esto inferir el valor de algún punto particular de interés.

Otra dimensión a considerar se ve reflejada con el patrón que definimos como *Búsqueda de una señal*, donde las respuestas arrojan que ante la dificultad que les representa la incertidumbre o la variación intrínseca en este tipo de problemas, los alumnos tratan de resolver la situación en un terreno familiar para ellos o con el que se sienten cómodos, posiblemente por esto utilizan en su mayoría procedimientos aritméticos (regla de tres, factor de proporcionalidad o sumas y restas) para hacer la estimación; donde además solo utilizan una parte de los datos. También es el caso del problema de la recta de ajuste, en el que ante la ausencia de la noción de incertidumbre o de esa

visión de agregado, utilizan parcialmente los datos disponibles, haciendo referencia al modelo de la función lineal como alternativa para ajustar una recta a una distribución de puntos que presenta una tendencia lineal, considerando sólo algunos de estos puntos en su elección, ignorando la influencia de la totalidad de estos y su variación conjunta, y sobre todo, definiendo dos tipos de datos: los que pertenecen o no al modelo lineal que ellos eligen trazar.

Como tercera dimensión proponemos, *Sentido de Variabilidad*, representada por el código que se definió inicialmente como *Cercanía*. En este patrón de razonamiento se incluyeron las respuestas que hacen referencia, aunque de manera somera, a la percepción de que existe cierta variación que subyace en este tipo de problemas y que debe considerarse al momento de proponer una recta de ajuste para un conjunto de puntos, tal y como lo hicieron los estudiantes al trazar y argumentar que la recta debe posicionarse lo más cerca posible de la mayoría de los puntos.

Confiamos en que la identificación de estos patrones de razonamiento como la manera en que los estudiantes llevan a cabo inferencias informales ante situaciones-problema de asociación estadística abona al cuerpo del conocimiento en el estudio de datos bivariados, sin dejar de lado que es importante explorar obstáculos de aprendizaje como son la aparente desvinculación que tiene el estudiante entre la naturaleza predictiva o inferencial inherente al modelo de regresión lineal y su identificación como aquella recta que mejor se ajusta al conjunto de puntos, así como la dificultad para concebir al conjunto de datos como un agregado, es decir, como un sistema en el que están ligados unos a otros y tienen la propiedad de ser desviaciones de un mismo modelo.

Referencias

Bakker, A., Biehler, R., and Konold, C. (2004). Should Young Students Learn About Box Plots? In G. Burrill and M. Camden (Eds.), *Proceedings of IASE 2004 Roundtable on Curricular Development in Statistics Education*, Lund, Sweden. Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D., and Garfield, J. B. (Eds.) (2004). *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Ben-Zvi, D. (2006). Scaffolding students informal inference and argumentation. En A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education*: Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Brazil. Descargado de http://www.stat.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf

Casey, S. (2014). Teachers´ knowledge of students´ conceptions and their development when learning linear regression, en K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics. ICOTS9 Invited Paper-Refereed*. Voorburg, The Netherlands: International Statistical Institute

Casey, S., & Wasserman, N. (2015), "Teachers' Knowledge about Informal Line of Best Fit."*Statistics Education Research Journal*, *14*(1), 8-30.

Moore, D. S. (1988). Should mathematicians teach statistics? *The College Mathematics Journal*, *19*(1), 3–7.

Moritz, J. B. (2004). Reasoning about covariation. In D. Ben-Zvi, & J. Garfield (Eds.) *The challenge of developing statistical literacy, reasoning and thinking* (pp.227-256). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pfannkuch, M. (2005). Informal inferential reasoning: A case study. In K Makar (Ed.), *Proceedings of the International Forum for Research in Statistical Reasoning, Thinking and Literacy*. Auckland, NZ. Brisbane: University of Queensland.

Rubin, A., Hammerman, J., and Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman and B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. Voorburg: The Netherlands: International Statistical Institute.

Zieffler, A. S., & Garfield, J. (2009). Modeling the Growth of Students' Covariational Reasoning during an introductory statistics course. *Statistics Education Research Journal, 8*(1), 7-3.