THE INFORMAL COVARIATIONAL STATISTICAL REASONING: FOCUS ON THE NOTION OF AGGREGATE USING DIGITAL TECHNOLOGY

Cindy Nathalia Morgado Hernández
Center for Research and Advanced Studies (Cinvestav), Mexico
cindy.morgado@cinvestav.mx

Ernesto Sánchez Sánchez
Center for Research and Advanced Studies (Cinvestav), Mexico
esanchez@cinvestav.mx

We report the results of a study on informal covariate statistical reasoning conducted with 22 students (aged 16 and 18 years). We designed and implemented a task in a digital technology environment to introduce the line of best fit. The task design having elements that foresee misconceptions reported in the literature, and by focusing on four statistical ideas that we consider being central to the development of informal reasoning about the line of best fit. After having used the digital technology environment, students transitioned from viewing points of a scatterplot as individual points or fragmented into subsets to viewing the scatterplot as an aggregate from a mathematical mechanism that links them through the notion of distance from a point set to a right line.

Keywords: informal covariational reasoning, line of best fit, aggregate, digital technology

Introduction

In statistics, covariation is the variation of two statistical variables that take numerical values (Moritz, 2004).The values for each variable are obtained from the same observation unit and expressed as an ordered pair; observations compose a set of pairs called bivariate data. The graphical tool used to represent a bivariate data set in a plane is a scatterplot. The most used techniques to investigate statistical covariation are correlation and regression. Correlation quantifies the strength of the linear relationship between a pair of variables, while regression expresses the relationship as a mathematical model (linear equation).

The correlation and regression are statistical objects that express global properties of a data set, properties that do not belong to isolated points of the data set but all of them. An isolated individual data does not contain the properties that will emerge when a data is associated with other data, namely the data viewing as an *aggregate*. Stigler (2016) used the term aggregation to designate the first pillar of statistical wisdom. For him, aggregation is the mechanism by which can provide more information of a data set, there is a loss of information of the individual data for retaining global properties of the data set.

Statistical educators have mentioned the aggregate to highlight a recurrent phenomenon in the learning of statistical concepts. Hancock et al. (1992) were the first to raise the problem that in data analysis, students are prone to focus on the characteristics of individual data without making sense of the aggregate properties of a data set, such as the mean. Other authors have mentioned the same problem concerning the notion of distribution (Bakker & Gravemeijer, 2004), in group comparison (Ben-Zvi & Arcavi, 2001), and the concept of the sample (Saldanha & Thompson, 2002). The correlation and regression result from an aggregation process and require that a point could be conceived of as an aggregate.

With the availability of specialized statistical applications and software, opportunities open up in teaching for students to look at data sets as an aggregate; the possibility of multiple representations, dynamic trawling, real-time data updating, and performing tedious calculations are ideal to show how statistical summaries are related to the data set from which they originate (Biehler et al., 2013). This paper informs about a design research that use the online data analysis platform CODAP, and an applet designed in GeoGebra with the aim to promote the development of covariational reasoning in high school students; In particular, we are interested in observing if the design of the task and the use of the software allow students to begin to see the points of scatterplot as an aggregate or, more precisely, the process of finding the line of best fit as an aggregation process.

Background

Many researchers have been interested in the problems of the correlation and regression teaching and learning. The first studies were about the conceptions of university students, such as of Truran (1995), who was interested in the detection and characterization of the interpretations of university students about the correlation coefficient and the determination coefficient. Sánchez-Cobo et al. (2000) studied verbal, graphical, and numerical representations of correlation, and Sorto et al. (2011) studied students' conceptions of the line of best fit. At the pre-university school levels, there is the study by Watson and Moritz (2007) analyzed students' reasoning when making graphical representations about covariation, and Casey (2015, 2014) studies of students' conceptions of what the line of best fit is. Regarding high school students' conceptions, two studies related to the graphical representation of covariation stand out: the study by Watson and Moritz (1997) analyzed the graphical representation established by students about the covariation present in non-symbolic contexts, and the study of Estepa and Batanero (1996) established some conceptions of covariation in students when they judge the relationship of two variables based on scatterplots.

A current trend in statistical education is to conduct research that investigates the relationships between teaching design and progress in student learning. In this way, studies have begun in which the design of the intervention in the classroom is an important component highlighting technology as an element that can help students make more accurate covariation judgments (Batanero et al., 1998; Cobb et al., 2003; Inzunza, 2016). At the high school level, research that includes teaching intervention is still scarce and scattered; we found only three studies at this level, each paper covering one topic: covariation in big data contexts (Gil & Gibbs, 2016), scatterplots, and the line of best fit (Medina et al., 2019), visualization and trend in the data (Dierdorp et al., 2011).

Framework Conceptual

The following four subsections present the concepts we consider central for understanding the research from which this report was done. We defined the conceptual framework used here as a set of concepts that clarify the key aspects to be studied. We aligned this notion of the conceptual framework with the one presented by Miles and Huberman (1994, p.18).

Definition of Aggregate and Aggregation in Statistics

From examining how researchers use the term aggregate, we define it as a set of data belonging to a larger whole with global properties. An aggregation process comprises an object produced from an aggregate that highlights properties common to all data, properties that individual data does not have. Thus, the mean and the line of best fit result from an aggregation process. A condition for the mean of a data set or the line of best fit of a "cloud"[1] of points to be seen as representatives of their respective data sets is that we must conceive them as aggregates. This characterization tries to synthesize the comments of Hancock et al. (1992), Konold and

Higgins (2002), Bakker and Gravemeijer (2004), Casey (2015), and Stigler (2016). From an educational perspective, the question arises: How does a student come to conceive that a data set is an aggregate? We hypothesize that there are two sources: context and mathematics. We can perceive that all the data belong to a phenomenon or associate them with a causal relationship. Here, we see the data within a context that unifies them. Also, we can perceive the data as part of an aggregate if it belongs to a set linked by a mathematical property, for example, even numbers in a set of numbers, cloud points aligned in a bivariate data set, etc. In both cases, the totality transcends the given set such that we can imagine the existence of other data that could be added to the given set.

Beliefs, Conceptions and Difficulties about Covariation

The literature reports that students do not separate their previous beliefs for observing and evaluating the behavior of two quantitative variables, and they do not include the word variation in their vocabulary (Moritz, 2004). Also, the previous concepts of a linear function in mathematics can interfere in their ability to make sense of determining the line of best fit (Casey & Nagle, 2016). Estepa and Batanero (1996) established some conceptions in the students when they evaluate the covariation and the line of best fit. Deterministic conception when students considered the relationship between the variables from a functional point of view (a line that passes through all the points), they expect a correspondence where each value of the dependent variable correspond to another value of the independent variable when this is not the case, consider that there is no dependence between the variables, local conception when they use only a part of the data and they generalize conclusions to the entire data set. Casey (2015, 2014) establishes the following strategies for students to draw the line of best fit: draw a line that divides the data points so that half of the points are at the top of the line and half are below the line, draw the line through the midpoints of different cloud groups. In addition, in the students may emerge the concept of "closeness" between the line and the point cloud but there is a lack of understanding of other elements such as the error that corresponds to the sum of the squares of the vertical distances between the observed and predicted values.

Informal Covariational Statistical Reasoning

Reasoning refers to the processes of obtaining and verifying propositions (conclusions) based on evidence or established knowledge or assumptions. Reasoning can take many forms, ranging from informal argumentation to deductive demonstration (National Council of Teachers of Mathematics, 2009, p.5). Informal statistical reasoning is related to data, samples, chance, inference, and relationships between statistical variables. Informal statistical reasoning about covariation is related to bivariate data sets and relationships between statistical variables. In the present research, the purpose is to develop students' informal statistical reasoning about the line of best fit. For this purpose, we define informal notions of linear (instead of quadratic) distance and line of best fit, which does not coincide with the formal concepts but is not inconsistent with them and has the advantage of being closer to students' intuitions.

The Influence of Digital Technology on Reasoning about Statistical Covariation

Research suggests that technology can help students make more accurate covariation judgments (Batanero et al., 1998; Cobb et al., 2003; Inzunza, 2016). Technology plays a very important role in statistics since it makes them visual, interactive and dynamic, allowing a focus more on concepts rather than algorithms and calculations, where interactivity and the quality of use of graphs allow conducting experimentations with data; this allows engaging students in productive activities (Biehler et al., 2013). Specifically, the topics of regression and correlation with the technology possess the following relevant features: 1) The possibility to form

Olanoff, D., Johnson, K., & Spitzer, S. (2021). *Proceedings of the forty-third annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Philadelphia, PA.

scatterplots and fit a regression line by visually showing the changing quadratic deviations of the line as it fits the cloud of points. 2) Obtain the numerical value of the correlation coefficient and determine the algebraic expression of the regression line. 3) Dragging points from the scatterplot and observing in real-time the effect of their location within the cloud on the strength (correlation coefficient) and direction (regression line) of the relationship, allowing one to see the interactions between the elements dragged and the statistical measures. 4) Linking multiple representations to discover and observe patterns and trends in data simultaneously from different perspectives (the graph, the summary measures, the regression line). In the present study, we show how with the help of technology, students can conceive of the line of best fit as the result of an aggregation process.

Methodology

The study participants were 22 high school students around the age of 17 who had not studied the topics related to correlation and linear regression. The application of the task took place in a computer classroom during a two-hour class session. The author of the present work carried out the implementation of the task. The data obtained were the worksheets developed by pairs of students.

We follow the principles of the design experiment of Cobb & McClain (2004) for the design and implementation of the task: the *use of technology*, we used *CODAP* and *GeoGebra* software because their features allow designing elements that we consider relevant for students to observe and interact with. For the *classroom discourse*, the teacher oversaw, monitoring, coordinating, and making sense the interaction between the students and between students and the technological tool. The *structure of the task* in the classroom is collaborative work students had the opportunity to explore solutions, compare them with those of their peers, and clarify them in a group meeting. To establish the *central statistical ideas*, we analyzed the difficulties and conceptions reported in the literature to promote elements of the task that anticipate the difficulties and students' conceptions. Also, we reviewed the content stipulated by the NCTM (2000, p.327-328) for the study of bivariate data (correlation and regression) in the last school level of high school and the bivariate data unit of the program of the College of Sciences and Humanities of the UNAM, Mexico.

The task starts from the graphical view in GeoGebra, where the data in Table 1 are displayed. The data correspond to the measure in fat gain (in kilograms) and change in energy use (in calories) from other "non-exercise activity" (NEA) (restlessness, daily life, and the like) of 12 young adults who overfed for eight weeks. We suggest that the teacher conduct a discussion asking whether changes in restlessness and other non-exercise activities explain weight gain in overeaters, guiding students to study the relationship between the variables on pencil and paper, followed by making the scatter plot in *CODAP* and leading them to describe the behavior of the point cloud (intensity and direction).

Table 1: Measures of Change in NEA and Fat Gain in the 12 Young Adults

NEA change (cal)	-94	-57	-29	135	143	245	355	486	535	571	620	690
Fat Gain (kg)	4.2	3	3.7	2.7	3.2	2.4	1.3	1.6	2.2	1	2.3	1.1

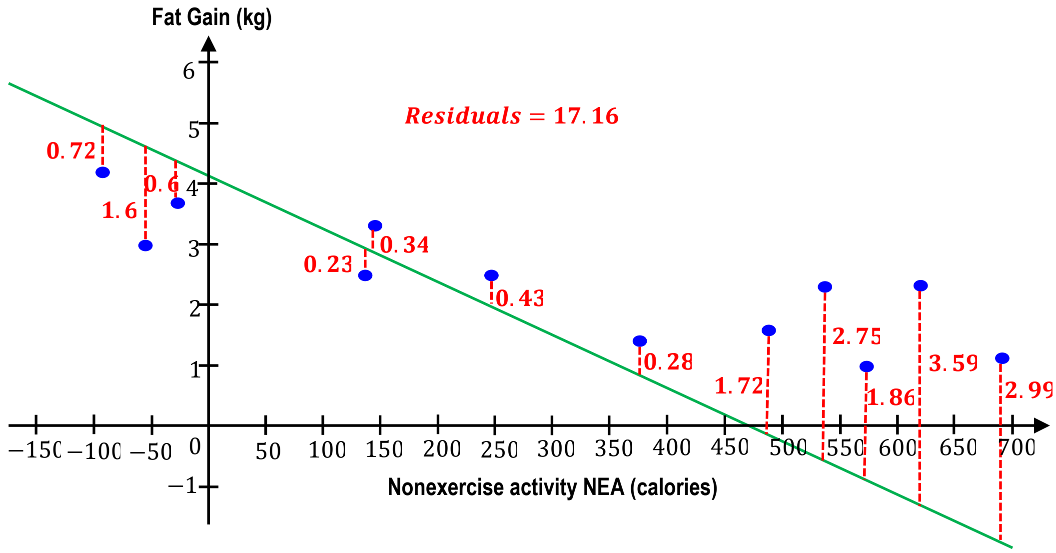

Figure 1: Screenshot of the Graphical View of the Task in GeoGebra

The *central statistical ideas* related to the line of best fit are present in the elements that make up the task and are:

1. The possible line always depends on all points in the scatterplot, i.e., all points influence the determination of how close or far away a line is.

The graphical view of GeoGebra shows a line that is movable by holding the click anywhere on it this line moves in different ways as desired, varying its slope or varying its point of intersection with the axis. The intrinsic characteristic of the movable line of being connected to the points (data in Table 1) focuses on the idea of conceiving a point cloud as an *aggregate*, i.e., all points in the cloud influence the determination of the best fit line.

2. Given a cloud of points and a line, we define the error of a pair of points (called residual) as the absolute value of the difference between the ordinate of the point and the ordinate of the projection of the point on the line of fit.

For each dashed segment in the graphical view, we show the numerical value called residual that corresponds to the difference between the ordinate of each data and the ordinate of the point that belongs to the moving line. What happens to the value of each residual if you move the line near or far from the point cloud? We intend to focus on the statistical idea of error, seen as a distance between a moving line point and a cloud point (datum).

3. Adding all the residuals, we obtain the measure (residuals) of closeness or remoteness between the cloud of points and the moving line.

In the file, there is a value called residuals. The teacher should mention that the residuals correspond to the sum of all residuals, and we calculate them by adding the absolute values of the difference between the points of the cloud and the possible points of the movable line. The objective is to provide a notion of distance from a line to a point cloud, and with this, to define a measure of the closeness of the line to the point cloud; this distance is the sum of all residuals. The student will explore how the distance from the line to a cloud changes by freely moving the

Olanoff, D., Johnson, K., & Spitzer, S. (2021). *Proceedings of the forty-third annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Philadelphia, PA.

line and observing the corresponding distance value. We suggest asking the students: Where do you think you should place the line so that the value of the residuals is the minimum? Can you be sure that the location you consider is unique, yes, or no? why?

4. The line that best fits the data is the one where the sum of residuals is minimal.

With the idea of defining the line that best fits the data as the one where the errors (residuals) or vertical distances are the smallest possible in some average sense in general, we suggest that the teacher indicate how to determine the line of best fit that GeoGebra yields and ask Is the way you placed the movable line the same as the line yielded by GeoGebra, yes, or no? How do they differ? What do you think is the criterion that GeoGebra uses to determine the line of best fit? We use a notion of linear distance and not quadratic distance because the former is more intuitive for the students., however, it seems clear to us that, understanding the idea with linear distance, it can be easily generalized considering the quadratic distance, arguing the reason for the advantages of this one.

Findings & Discussion

The analysis of the students' responses in judging the relationship between the variables shown in the problem situation together with the table of values without the use of technology provides the following reasoning as a result:

The *functional covariation* strategy comprises searching for and isolating bivariate data that adhere to a mathematical model, i.e., they focus their attention on the points they locate on a line. Students whose solutions fall into this category divide the data set into two parts, those that correspond to a linear model and those that do not. So their description of the data refers only to the subset of data that corresponds to the model and ignores those that are left out, and based on the data they selected, they describe the general trend, but their statements, often coinciding with the trend of the entire cloud, state, for example: "the more fat you consume, the more calories you increase", "the less you change in calories through movement, the less fat you will decrease". They also involve in their description's characteristics of the context of the problem*:* "fat increases because it remains encapsulated", "the less activity you do, the more fat", "the fats remain in the body". Their beliefs may influence their choice of the point cloud data set other than following the criterion that they lie on a straight line.

When students focus their attention on what happens in the passage from one point to the next within the values table, they present the strategy of *randomness.* They tracked the differences between successive points as they review the points from left to right, and they do not notice any predictable patterns. Sometimes, the difference is positive or negative, and the size varies. Thus, they conclude: "the fat increases or decreases depending on how many calories are burned in ANE", "if the calories decrease or increase depending on the ANE, the fat will also increase or decrease", "the calories that are burned will increase the kg or calories that are burned will decrease the kg", the students who follow this strategy, it is evident that there is no correlation since it is not possible to know if from one value to another value fat levels will go up or down. In both strategies, they visualize the data by paying attention to their individual or partial characteristics, and not as a whole. Thus, the analysis according to the first strategy consists of separating the data into two, in the second, of going through them one by one and seeing if and by how much they increase or decrease. When they use the CODAP platform to enter the table of values and make the scatter diagram and again to judge the relationship between the variables, the *inverse covariation* arises, which globally describes the behavior of the cloud, for example: “when increasing calories there is more fat-burning”, “the more change in ANE, the less fat increase there is”, “when the change in ANE is greater, the increase in fat is less”, “the more increase in ANE, the more fat loss there is. The less increase in ANE, the less fat loss”, the *context of the problem* also influences, for example: “the fat does not increase because, although you do not exercise if you keep moving”, “the increase in fat depends on the amount of ANE that is carried out daily”. Without using technology, students focus their attention on the variation from point to point (Randomness: fat sometimes increases and sometimes decreases), while that using the technology influences to make a description of the general trend the cloud, ignoring particular fluctuations (*inverse covariation:* fat values decrease as ANE values increase).

We see that treating the situation with the help of technology and the use of some of its possibilities influences students to abandon their tendency to highlight in their description of the cloud to a set of individual points that they compare with each other, to highlight a global property which becomes evident when you have an aggregate view of the set of data. In addition, with and without technology, the description that they did of the relationship between variables is influenced by the contextual content of the problem. Students do not see the data as simply numbers, but numbers with a context, which for Moore (1990) is what it implies to establish the statistical association.

Regarding the analysis of the students' responses when asked to establish the line of best fit to the data using the GeoGebra applet, the following reasoning emerged:

In many student responses when using technology, the intuition of the *closeness* of the set of points to a line is revealed; for example: "that the blue points were closer to each other to the line", "I think it is the line that passes near all the points and does not join them, so each one has a certain distance". This intuitive idea becomes operative with a notion of *distance* from a line to a set of points. The software allows to calculate the residuals, that is, the differences of the ordinates of each pair of points with the same abscissa, one belonging to the cloud and the other on the line. Each residual can be viewed as a distance from the point of the cloud to the line (for the present purpose it does not affect the fact that strictly the distance from a point to a line is defined as the distance of the segment that passes through the point perpendicular to the line) and the sum of residuals as the total distance of all residuals from the cloud to the line. As this sum can be seen in the software and is updated in real-time as the line moves, the students manipulate and see a real function determined by the point cloud and whose independent variable is the lines in the plane; then for them, the line that minimizes the function is the line closest to the points, that is, the one with the best fit. The following expressions of the students indicate some of their ideas in this regard: “the less distance there is between the points and the line, the remainder will change”, “the further the line of adjustment of the points in the table is in the graph is greater the value of residuals since there is a greater distance", " when the line is better centered, the level of residuals is lower, that is, it passes centering in the middle of the points.

One condition for viewing a set of data as an aggregate is to imagine a rational mechanism that unites them all into a totality that represents them, even if individual information is lost. When students choose a subset on a line from the point cloud, they consider the subset as an aggregate but ignore and discard some data; they do not see the cloud as an aggregate. The process of looking for the line closest to the point cloud with a notion of distance shows that each one of the points contributes to determining said line. This fact is significant because the literature has highlighted that one problem for students to understand the statistical notions of centers, variation, data comparison, and data distribution, is that they see the budding data set as an added (Bakker & Gravemeijer, 2004; Konold & Higgins, 2002; Roseth et al., 2008). Also, the context is an important factor to see a data set as an aggregate since in this case, it is the existence of a *causal mechanism* (the term is from Zimmerman, 2007) that suggests that all the data are related, in our case, that ANE and fat accumulation are part of a causal process.

Conclusions

A necessary condition for understanding the topics of regression and correlation is conceiving a set of bivariate data as an aggregate. However, students do not reach this conception spontaneously and probably not with traditional methods as they have been taught the topics at the high school level. There are two levels in which it is convenient to analyze the appearance and development of a set of points as an aggregate: the mathematical level and the contextual level. At the mathematical level, technological resources allow students to move from seeing a cloud of points as individual points or fragmented into subsets to seeing it as an aggregate based on a mathematical mechanism that links them with the notion of distance from a cloud to a straight line. At the context level, it is the existence of a causal mechanism that helps to imagine a unity in the data set. It would remain to work on the sources of variability for students to explain why the data differ from the probable causal model.

Note

[1] We use the term "cloud" of points or simply "cloud" to refer to bivariate data represented in the scatterplot.

References

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

Batanero, C., Godino, J. D., & Estepa, A. (1998). Building the meaning of statistical association through data analysis activities. In A. Olivier & K. Newstead (Eds.), *Proceedings of the 22 Conference of the International Group for the Psychology of Mathematics Education* (pp. 221–236). University of Stellembosch.

Ben-Zvi, D, & Arcavi, A. (2001). Junior High School Students' Construction of Global Views of Data and Data Representations. *Educational Studies in Mathematics*, *45*(1/3), 35–65. http://www.jstor.org/stable/3483096

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clement, J. Bishop, C. Keitel, J. Kilpatrick, & A. Y. L. Leung (Eds.), *Third International Handbook of Mathematics Education* (pp. 643–689). Springer. https://doi.org/10.1007/978-1-4614-4684-2_21

Casey, S. A. (2015). Examining Student Conceptions of Covariation: A Focus on the Line of Best Fit. *Journal of Statistics Education*, *23*(1), 1–33. https://doi.org/10.1080/10691898.2015.11889722

Casey, S. A. (2014). Teacher's knowledge of students' conceptions and their development when learning linear regression. In K. Makar, B. Sousa, & R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July 2014)* (pp. 1–6). International Statistical Institute.

Casey, S. A., & Nagle, C. (2016). Students' use of slope conceptualizations when reasoning about the line of best fit. *Educational Studies in Mathematics*, *92*(2), 163–177. https://doi.org/10.1007/s10649-015-9679-y

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design Experiments in Educational Research. *Educational Researcher*, *32*(1), 9–13. http://www.jstor.org/stable/3699928

Cobb, P., & McClain, K. (2004). Principles of Instructional Design for Supporting the Development of Students' Statistical Reasoning. In Dani Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 375–395). Springer Netherlands. https://doi.org/10.1007/1-4020-2278-6_16

Dierdorp, A., Bakker, A., Eijkelhof, H., & van Maanen, J. (2011). Authentic Practices as Contexts for Learning to Draw Inferences Beyond Correlated Data. *Mathematical Thinking and Learning*, *13*(1–2), 132–151. https://doi.org/10.1080/10986065.2011.538294

Estepa, A., & Batanero, C. (1996). Judgments of correlation in scatterplots: Students' intuitive strategies and preconceptions. *Hiroshima Journal of Mathematics Education*, *4*, 24–41.

Gil, E., & Gibbs, A. (2016). Introducing secondary school students to Big Data and its social impact: A study within an innovative learning environment. In J. Engel (Ed.), *Promoting understanding of statistics about society. Proceedings of the Roundtable Conference of the International Association for Statistics Education (IASE)*.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic Inquiry With Data: Critical Barriers to Classroom Implementation. *Educational Psychologist*, *27*(3), 337–364. https://doi.org/10.1207/s15326985ep2703_5

Inzunza, S. (2016). Análisis de datos bivariados en un ambiente basado en applets y software dinámico. *Revista Educación Matemática*, *28*(3), 61–89.

Konold, C., & Higgins, T. L. (2002). Highlights of related research. In & V. B. S.J. Russell, D. Schifter (Ed.), *Developing mathematical ideas: Working with data* (pp. 165–201). Dale Seymour Publications. https://srri.umass.edu/publications/konold-2002hrr/

Medina, M., Sanchez, E., & Silvestre, E. (2019). Covariational reasoning patterns exhibited by high school students during the implementation of a hypothetical learning trajectory. In Uffe Thomas Jankvist and Marja van den Heuvel-Panhuizen and Michiel Veldhuis (Ed.), *Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education (CERME11)*. https://hal.archives-ouvertes.fr/hal-02412811

Miles, M. B., & Huberman, A. M. (1994). Qualitative data analysis: An expanded sourcebook, 2nd ed. In *Qualitative data analysis: An expanded sourcebook, 2nd ed.* Sage Publications, Inc.

Moore, D. S. (1990). Uncertainty. In *On the Shoulders of Giants: New Approaches to Numeracy* (pp. 95–137). National Academy Press.

Moritz, J. (2004). Reasoning about Covariation. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 227–255). Springer Netherlands. https://doi.org/10.1007/1-4020-2278-6_10

National Council of Teachers of Mathematics. (2009). *Focus in High School Mathematics: Reasoning and Sense Making*. NCTM.

Roseth, C. J., Garfield, J. B., & Ben-Zvi, D. (2008). Collaboration in learning and teaching statistics. *Journal of Statistics Education*, *16*(1). https://doi.org/10.1080/10691898.2008.11889557

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*(3), 257–270. https://doi.org/10.1023/A:1023692604014

Sánchez-Cobo, F., Estepa, A., Batanero, C., Castro, E., & Bernabeu, B. (2000). Un estudio experimental de la estimación de la correlación a partir de diferentes representaciones. *Enseñanza de Las Ciencias*, *18*(2), 297–310.

Sorto, M. A., White, A., & Lesser, L. M. (2011). Understanding Student Attempts to Find a Line of Fit. *Teaching Statistics*, *33*(2), 49–52. https://doi.org/10.1111/j.1467-9639.2010.00458.x

Stigler, S. M. (2016). *The Seven Pillars of Statistical Wisdom.* Harvard University Press.

Truran, J. (1995). Some Undergraduates' Understanding of the Meaning of a Correlation Coefficient. In B. Atweh & S. Flavel (Ed.), *Annual conference; 18th, Mathematics Education Research Group of Australasia (MERGA)* (pp. 524–529). Northern Territory University.

Watson, J. M., & Moritz, J. (1997). Student analysis of variables in a media context. In B. Phillips (Ed.), *Proceedings of the 8th International Congress of Mathematics Education (ICME8)* (pp. 129–147). Swinburne PresS.

Watson, J., & Moritz, J. (2007). Developing aspects of distribution in response to a media-based statistical literacy task. *International Statistical Institute 56th Session*.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223. https://doi.org/https://doi.org/10.1016/j.dr.2006.12.001