Informal Inferential Reasoning on Simple Linear Regression Encouraging Critical Thinking

Marte Bråtalien[1*] and Margrethe Naalsund [1]

[1] Department of Educational Sciences, Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

*Corresponding author. E-mail: marte.bratalien@nmbu.no

Abstract: Informal inferential reasoning (IIR), described as making evidence-based generalizations about a population based on samples, is considered important for the development of argumentation-, inference-, critical thinking - and aggregate thinking abilities. This article aims to explore how undergraduate students' IIR can develop in an inquiry problem-solving session on simple linear regression, through individual think-aloud protocols with follow-up conversations with five students. Our findings suggest that enabling and supporting the students to grapple with their own hypotheses is important for their development of IIR encouraging critical thinking. From initial hypotheses with limited argumentation and little regard to the probabilistic nature of statistical inferences, the students' reasoning evolved in terms of making probabilistic generalizations from data when they were given time and concurrent probing to elaborate on, and question, their own arguments and inferences. We also suggest that in addition to emphasize the signal in the noise, giving attention to the noise around the signal can be fruitful for their IIR.

1 Introduction

An important development in research on statistical learning is the change in focus from procedural skills and methods towards reasoning and aggregate thinking (Ben-Zvi & Garfield, 2004; Makar & Rubin, 2009). This shift has brought increased attention to critical thinking, inquiry, and inference, which are considered important for in-depth learning and critical application of knowledge (e.g., Artigue & Blomhøj, 2013; Dorier & Maaß, 2020; Garfield & Ben-Zvi, 2008). Moreover, inference "is at the heart of statistics, as it provides a means to make substantial evidence-based claims under uncertainty when only partial data are available" (Makar & Rubin, 2018, p. 262), which makes learning and teaching about inferences essential for statistics education (Pratt & Ainley, 2008).

DOI: 10.5324/njsteme.v8i2.4943

While formal statistical inference is based on formal statistical procedures (Makar & Rubin, 2009; Zieffler et al., 2008), informal statistical inference involves reasoning based on critical thinking and knowledge without using formal statistical procedures. *Informal inferential reasoning* (IIR) concerns the drawing of evidence-based conclusions about a wider population from sample data, and has shown promising effects on developing students' understanding of statistical key concepts (Ben-Zvi, 2006; Konold & Pollatsek, 2002; Makar & Rubin, 2009; Zieffler et al., 2008). We view IIR as both the process of drawing and evaluating inferences, and the product of such processes, without trying to separate the two. Similar definitions are used in other research studies on reasoning (e.g., Lithner, 2008; Makar & Rubin, 2009).

Few researchers have investigated students' IIR on correlations, for example through scatterplots (Makar & Rubin, 2018). Thus, there is a need for research addressing IIR in this statistical context. One approach is to explore IIR on a *micro-level*, focusing on how students' informal reasoning develops over a defined period of time (Pratt & Ainley, 2008). In this light, the study presented in this article explores how five undergraduate students' IIR develop through a micro-level position in individual 30-minute inquiry problem-solving sessions on simple linear regression. Such research can provide detailed insight on for example what data in the scatterplot the students base their informal inferences on, the statistical and non-statistical arguments they use to support their inferences, and what might initiate and guide any developments in their IIR in a short-time perspective (for example over the course of one activity). Further, this insight might inform both practitioners and researchers in design and orchestration of learning activities that support students' IIR.

Recent Nordic research has looked at students' emphasized criteria when informally fitting a line to a scatterplot (Bråtalien & Naalsund, 2021) and students' thoughts on what mathematical ideas computers use to find the best fitted line (Petersson, 2022). The first study adds to the body of research on how informal curve-fitting often is based on uncomplicated and easily operationalizable ideas, but also reveals how students might combine multiple ideas in their reasoning. The second study – asking the students explicitly to discuss the mathematical ideas that curve-fitting could be based on – found similarities between several student suggestions and central mathematical ideas for curve-fitting throughout mathematics history. Both studies provide interesting insight into students' reasoning as they construct linear regression models, but neither bring to focus what inferences the students believe can be drawn from the models they make. Drawing inferences related to linear regression and being able to critically discuss generalized conclusions and predictions about the correlation and causation between two variables, is central in several areas of education, work life and research. In addition, the increased attention to contrasting perceptions of the truth in different media, often accompanied by statistical data or models as "evidence", makes it important to critically discuss and interpret sample data and the inferences that are drawn from them, for example regarding representativeness, sampling, correlation, and causation. Further, in our everyday life we make numerous conclusions, predictions and decisions based on the selected data that we have (Makar & Rubin, 2018), for example from experience or previous knowledge. In other words, we shape our future based on inferences drawn from limited data. These types of inferences are not necessarily – and in our daily life most likely not – based on formal statistical methods, which highlights the importance of critical thinking and informal inference. In this article, we explore the following research question: *How can undergraduate students' IIR on simple linear regression develop in an inquiry problem-solving session?*

2 Informal Inferential Reasoning

Makar and Rubin (2009) offer a theoretical framework on IIR in statistics, highlighting three key principles: *Generalizations beyond the data*, *the use of data as evidence* and *probabilistic language*. Similar principles are highlighted in other frameworks on IIR as well (e.g., Ben-Zvi et al., 2007; Zieffler et al., 2008). Based on this consensus, and the lucid explanations offered by Makar and Rubin (2009), we adopt their framework to interpret the students' IIR. Generalization refers to predictions, estimations and conclusions that go beyond describing the sample it is based on, often used to generate a hypothesis about a greater population or to evaluate it (Makar & Rubin, 2009). The generalizations must be based on some form of evidence, which can include observations, descriptions, numerical data and even unrecorded data (Makar & Rubin, 2009). When drawing and communicating inferences, the language used should be probabilistic, which means "any language appropriate to the situation and level of students to suggest uncertainty in a speculated hypothesis, that a prediction is only an estimate, or that a conclusion does not apply to all cases" (Makar & Rubin, 2009, p. 87). Makar and Rubin (2009) formulate the three key principles together as *probabilistic generalizations from data*, underpinning their intertwinement in inference processes. Even though the focus in IIR is on generalizations beyond the data, the data at hand serves a critical role. Being able to properly and critically evaluate the validity of claims based on the data, and the data themselves, is important for understanding statistical relationships (Ben-Zvi & Garfield, 2004). Moreover, generalizations should be accompanied by supporting arguments (Ben-Zvi, 2006; Ben-Zvi et al., 2007; Makar & Rubin, 2009; Zieffler et al., 2008), and data is required to build such arguments. This links generalizations beyond data to the use of data as evidence. Similarly, generalizing based on a data sample always includes some level of uncertainty. A different sample (e.g., other observations) would probably give a different scatterplot, which could lead us to draw completely different inferences. Any generalized inference must therefore be stated in probabilistic terms (Ben-Zvi et al., 2007; Makar & Rubin, 2009).

This study focuses on IIR in connection with simple linear regression, which involves making inferences by using samples to describe, interpret and model a possible general relationship between two variables. The sample can be illustrated as a scatterplot, which indicates whether there is an overall association between the response- and explanatory variable and the strength of the association between them. IIR is in our research used to describe the construction of generalizations, predictions, and conclusions, from mainly reasoning on a scatterplot. In this complex reasoning process of drawing informal statistical inferences, one must navigate through, and assess, great amounts of data to decide what evidence to build one's arguments on. Making inferences from scatterplots includes reasoning on sample size, variability, representativeness, signals and noise, interpretations, hypotheses, generalizations, uncertainties, and alternative explanations, all using a flexible view by shifting between emphasizing local points or clusters of data, viewing the data as an aggregate, and viewing the data as a fragment of a wider population. From these arguments, a generalization about the wider world forms – if the sample is considered representative for a larger population. "The key idea in inferential reasoning is that a sample provides some, but not complete, information about the population from which it is drawn" (Sotos et al., 2007, p. 101). This idea involves both the issue of representativeness, inasmuch as the characteristics of the sample should resemble the characteristics of the population (if the sample selection has been done properly), and variability between samples, inasmuch as different samples will reflect the populations to different extents (Batanero et al., 1994).

Research indicate that students might oversimplify statistical criteria for simple linear regression (e.g., Bråtalien & Naalsund, 2021) and use ideas and procedures that are only sufficient in special situations (e.g., Batanero et al., 1994). Students, across grade levels, often see sample data as a collection of single values or cases (Ben-Zvi, 2004; Ben-Zvi & Arcavi, 2001; Konold et al., 1997) in opposite to focusing on general patterns in the sample. Moreover, in their study on first year university students, Batanero et al.(1997) found that the students often based their conclusions and judgements on selected parts of the data. This can heavily affect their informal inferences, as the evidence for their inferences will be based on fragments of the sample (which again is just a fragment of a larger population). An aggregate view is considered fundamental to discover and explain patterns and draw inferences that take into consideration the variability in the data (Ben-Zvi & Arcavi, 2001), but several researchers argue that this can be a complex process (Ben-Zvi, 2004; Ben-Zvi & Arcavi, 2001; Konold et al., 1997). Some students believe that any sample, regardless of size, is representative (Sotos et al., 2007; Watson, 2004). This may lead to deterministic inferences, as the students do not reflect on how the sample affects the level of certainty their conclusions have. Students with such conceptions will typically generalize beyond what is appropriate. Another challenge is how students, even after formal instruction, struggle realizing that causation cannot be claimed solely from a strong association between two variables (Batanero et al., 1997). These findings are somewhat alarming, as generalizing and determining whether association in samples suggests casualization are key elements in statistical inference (Ben-Zvi, 2006).

3 Method

3.1 Participants and data collection

The data was collected in February 2017 at a Norwegian university. Five students participated: Emma, Heidi, Karoline, Linn, and Susanne. The study used purposive sampling in selecting participants (Bryman, 2016), and the five students were chosen in collaboration with the lecturer in a first course in university level elementary statistics, which all five students had participated in the previous autumn. The five students had all been verbally active in classroom discussions, which was considered positive for participation as verbalizing one's thinking is an important criterion for the chosen method (think-aloud protocol) to function optimally (Afflerbach & Johnston, 1984; Van Someren et al., 1994). Another important criterion was that they all freely agreed to participate in the study. We acknowledge that the all-female and talkative sample might not represent the average student in an elementary statistics course, which is not considered an issue, since the aim of the research was not generalizability to a population, but in-depth insight into how undergraduate students' IIR might develop throughout an inquiry problem-solving session.

The study was conducted approximately one month after the students had finished their final exam in the university level elementary statistics course, where linear regression was one of the topics taught. The reason for the delay between the course and the study was the university course schedule and holidays. The critical reader might question if we can talk about *informal* inferential reasoning when the students had received formal instruction on linear regression. Our view on this is that informal reasoning does not require that the students do not have any (formal) knowledge on the topic, only that the reasoning is done without using formal statistical procedures (for example, calculating R^2-values to evaluate the fit of a regression line would be a formal statistical procedure). Although we cannot rule out that the students' reasoning in this study might have been affected by their formal instruction on linear regression, the problem that the students worked on did not offer them information needed to do formal procedures (e.g., numbers from which they could calculate R^2-values) and, hence, the problem encouraged IIR. We outline the problem in the next subsection (3.2), while our actions to promote informal reasoning are further addressed in subsection 4.2.

Individual think-aloud protocols were used, which means that the students were encouraged to continually speak aloud their thoughts as they worked through a problem (Afflerbach & Johnston, 1984; Van Someren et al., 1994; Young, 2005). The think-aloud sessions were followed by a short conversation (cf. Van Someren et al., 1994) between the student and one of the researchers to enable further exploration of relevant reasoning sequences without interrupting the students' chain of thought during the problem solving. Thus, the researcher's role was initially to mainly provide prompts when the student was quiet for some time and to identify reasoning sequences to pursue in the conversations. Each think-aloud protocol lasted 30 – 45 minutes, including the conversation, and was video recorded.

There are some issues of reactivity connected to think-aloud data (Young, 2005), especially concerns related to the ability to think and attend to a problem at the same time, the effects of having to talk during problem solving, and the effect of drawing the attention to underlying cognitive processes whilst working on a problem. Measures were taken to ensure a calm and safe environment for the sessions: Research ethics were specifically discussed with the students, the interactions with the students during the sessions were limited, although encouraging and supportive, and the students were offered lunch after the sessions. There might be variations regarding how used the students are to verbalizing their thoughts, thus there is a risk for underestimating the students' reasoning abilities. However, this study does not aim to test students' abilities, nor classify them into different reasoning types, but to uncover possible developments in their IIR working with one data sample during a defined period. A central question then, is to what extent the think-aloud data reflects the students' reasoning. We believe that the post-activity conversation helped bring actions to consciousness that perhaps were unconscious during the activity, thus allowing us to gain further insights into the students' reasoning.

3.2 The problem

Tasks used to study IIR should challenge the students to build on their prior knowledge, draw inferences (without using formal statistical methods), and identify evidence to support their claims (Zieffler et al., 2008). The students were given the scatterplot in Figure 1, made from fabricated data pairs, with a line going through the plot. They were asked to discuss the model and what inferences they could draw from it (*"what can you say based on this model?"*). The line through the plot was *not* the best-fitted line. The reason for this was that the students later were asked to draw what they believed was the best-fitted line for the plot in Figure 1, without any calculations or technology to support them, and argue for the placement of their line. The students were then given the same plot but without the line. We have addressed their emphasized criteria for informally fitting the line elsewhere (Bråtalien & Naalsund, 2021); in this study we do not focus on their emphasized criteria but rather on their informal inferences they drew from the plot. The line seen in Figure 1 was placed in the plot to provide the students a starting point for their problem-solving and to encourage informal reasoning on for example variability, skewness, signals and noise, correlation, and causation. Formal measures such as the R^2-value was intentionally left out, and the placement and number of points discouraged the students from making such formal calculations themselves. The context, exam result versus amount of candy eaten, facilitated discussion on correlation and causation, in addition to relating the problem to the students' real world (as most students can relate to eating candy and studying for exams). We made a pedagogical decision to work with a sample of 19 datapoints, although the low sample size is insufficient for drawing conclusions. The reason for this was to highlight issues of low sample sizes, uncertainties, and representativeness, in addition to operationalize informal strategies for, and encourage discussion on, the placement of a regression line (as addressed in Bråtalien and Naalsund (2021)).

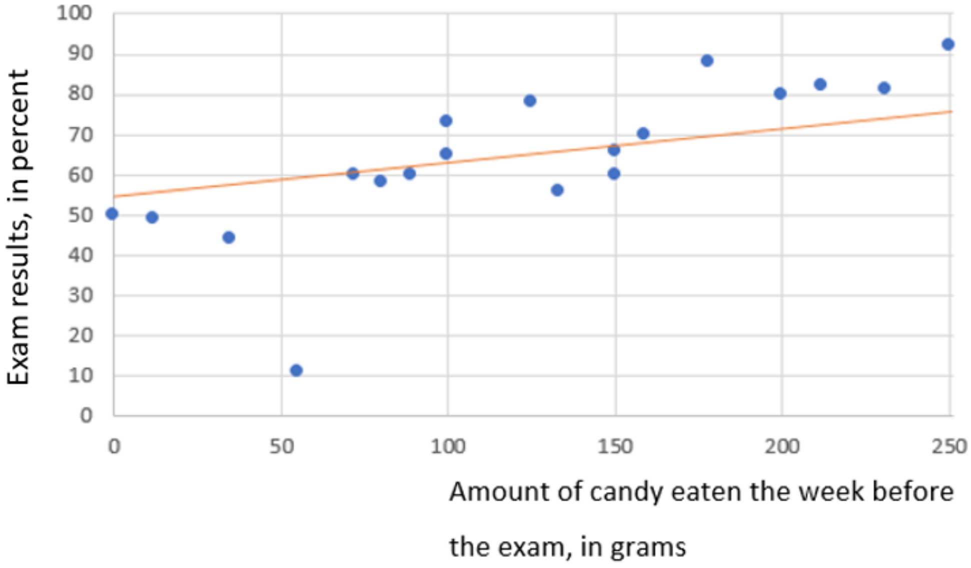

Figure 1. *The problem, a scatterplot with a line through it (not the regression line).*

3.3 Data analysis

The video recordings of the think-aloud sessions and subsequent conversations were transcribed to make easier the further process of coding and analysis. Through the transcription phase, we gained familiarity with the data (Lapadat, 2000; Powell et al., 2003). Thereafter, based on the theoretical framework and previous research, interpretive marginal notes were made alongside the transcriptions for each of the five students. "Coding and categorizing in this manner allow the researcher to identify the types of thinking evident in the think-aloud data" (Young, 2005, p. 26). The students' IIR was coded according to the three key principles of IIR: 1) making generalizations (e.g. hypotheses, predictions, conclusions that go beyond describing the sample), 2) the use of data as evidence (e.g. observations, variability, signals and noise), and 3) the probabilistic nature of the language used (e.g. reflection of uncertainty in evidence, hypotheses and conclusions).This structured the data into three overarching codes (generalizing, evidence, probabilistic language), each containing a substantial number of sequences from the transcripts. Both researchers discussed the codes, the coding procedures, and the interpretations.

The next analytical step was to write rich descriptions of each student's IIR working on the problem, from the coded sequences (Powell et al., 2003). The descriptions followed the development of their IIR through the session and a pattern (Stake, 2003) of three phases of the students' IIR grew forth, connected to the presence (or absence) of the key principles and any interconnection between them. The three phases are illustrated in Figure 2 and guide the results and discussion. Quotes are included that illustrate typical student IIR. The typicality is explained where the quotes are included. Thereafter, we identified reoccurring reasoning across cases, and differences between cases, that could offer useful insight considering our research aim. The similarities and differences were centered around the three key elements for IIR.

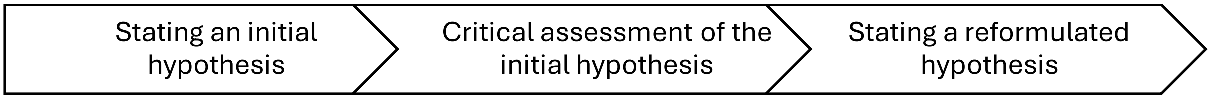
Stating an initial hypothesis
Critical assessment of the initial hypothesis
Stating a reformulated hypothesis

***Figure 2**. The five students' IIR process.*

4 Results and Discussion

All the students followed a similar process in their IIR (see Figure 2), even though they solved the problem individually. When handed the problem, they all looked within the data, before *stating an initial hypothesis* about the wider world. This was done within few minutes and with little to no interaction with the researcher. Building on this, the students, on their own or through inquiring interaction with the researcher, *critically assessed their initial hypothesis* through elaborating on the arguments it was based on and including new evidence and arguments – a process leading them to *state a reformulated hypothesis*.

4.1 Stating an initial hypothesis

The students all started their IIR by stating an initial hypothesis within just minutes after receiving the problem. Common for all five students was that this process was initiated by commenting the context (given by the axis labels) and the variability and placement of data points on the x- and y-axis (mainly focusing on skewness). The following extracts illustrate this, showing Emma and Susanne's individual first verbalized thoughts when receiving the problem. While Emma emphasized some selected "clumps" of data, Susanne seemed to focus on single data points.

Emma: (Receives the problem). So, I'm just trying to see where there are more [data points], and what [the scatterplot] really shows me. First of all, there are many students eating a lot of candy, it seems. (...) It tells me that the students with higher exam results eat a little more candy. But, then you have quite a few on the middle here, if I'm getting this right... So, in a way, the students with the best exam results have also eaten the most candy the week before their exam. So, from this, candy helps a little – one can assume that candy improves how you perform on the exam.

Susanne: (Reading the problem out loud). Mm... I see that on the x-axis we have amount of candy eaten the week before [the exam]. And here we have the exam result, and then it's plotted together. (...) If you look here (pointing) it's 150 grams of candy eaten, and that person got 60 percent correct on the exam, I guess. Well, that's just one person, there's many different. But it looks like the exam result actually improves, with the more candy you eat. (...) 'Cause it increases, and the more candy, the better the exam results get. So maybe it helps to eat candy the week before the exam, then.

Four of the five students formed initial hypotheses like Emma and Susanne's (Table 1), leading in the direction of causation in the form of candy intake improving the exam results. The last student (Linn) emphasized alternative explanations for the correlation between candy intake and exam results in her initial hypothesis. Common for all five students' initial hypotheses is that arguments and generalizations seemed to be based on parts of the data (Batanero et al., 1997), treating the sample as representative for a greater population (Sotos et al., 2007; Watson, 2004), and (apart from Linn) uncritically claiming a causal relationship (Batanero et al., 1997), all with little probabilistic language.

***Table 1**. The five students' initial hypotheses.*

Student	**Initial hypothesis**
Emma	*One can assume that candy improves how one performs on the exam.*
Heidi	*The more candy you eat, the better mark you'll get apparently.*
Karoline	*It seems like the more candy you eat, the better you do at the exam.*
Susanne	*It looks like the exam result actually improves, with the more candy you eat. (...) So, maybe it helps to eat candy the week before the exam, then.*
Linn	*Students who read a lot eat a lot of candy, and they also do well on their exams.*

We see the initial hypotheses that were made as what Makar and Rubin (2009) describe as generative hypotheses: “Speculative statements which are created by a reasoned process but for which their likelihood has not necessarily been systematically assessed” (Makar & Rubin, 2009, p. 86). As both Emma and Susanne’s transcripts illustrate, the “reasoned process” was short. The lack of justification of their claim of a causal relationship between candy intake and exam result supports Batanero et al. (1997) in that students might struggle, even after formal instruction, with the concept of causality. However, the students’ comments on the axis labels show that they were aware that the data did not exist in some vacuum but represented some sort of relation.

At first, we were puzzled by how the students’ arguments and emphasized evidence focused on a quick look at single values or clusters in the plot (clearly *within* the data) while the hypotheses were expressed as generalizations about a greater population (clearly *beyond* the data) with no arguments offered to justify this advance. Further, the inferences were expressed in rather non-probabilistic manners, especially considering the limited evidence. Nevertheless, formulating an initial hypothesis means creating a starting point. As Makar and Rubin (2009) point out and we will see in 4.2, this starting point can be followed by more critical assessment.

4.2 Critical assessment of the initial hypothesis

The following transcript follows directly from the previous transcript of Emma forming her initial hypothesis and shows the interaction between Emma and one of the researchers.

Researcher: Could you clarify what makes you say this [the initial hypothesis]?

Emma: Yes, well… You have exam results on the y-axis, and the dots that are the highest on the y-axis are also the furthest out on the x-axis, which means that they’ve eaten more candy. So, kind of, if you also look at the line that has been drawn to show the relationship here, it shows that there is a positive growth, which tells me that more candy also get better exam results. (…) There isn’t really many observations, though, so maybe this gives a bad representation of the reality. But, from the observations used here, it seems like this is the case.

Emma and Linn expressed a development in their IIR after none or just gentle probing from the researcher, through few and focusing questions, while Susanne, Heidi and Karoline’s reasoning developed through more substantial interactions with the researcher. The questions asked by the researcher were guided by each student’s arguments and hypothesis. The following transcript of the interaction between Susanne and the researcher after Susanne had stated her initial hypothesis, gives insight on the development in her IIR.

Researcher: What are your thoughts on what you just said [the initial hypothesis]?

Susanne: That it seems like a good idea to eat candy a week before the exam? (Laughs). Well, the students who ate little candy, like no candy, they had half of the exam right. And this is the regression line, I guess. The best fitted line for all the different single values here. In this sample, anyway.

Researcher: Could you add to that?

Susanne: Well, you have asked many different persons what they've eaten and what their exam results were, and then you get all these single values here. And you represent them like this, with a straight line that fits the best with all the values. (...) So, when you look at the distance from every value and to the line, then every value should be as close to the line as possible, in a way.

Researcher: All right. Could you say something about the fit, then?

Susanne: (...) [H]ere we see that many points are not on the line. So, the regression line maybe isn't really the truth. (...) Well, it could be better, so, that tells me that it might not always be true that eating candy the week before the exam gives a better exam result, but from this plot, it looks like this is the case.

Researcher: Mm... Is there anything that would have helped you say something about this relationship?

Susanne: Mm... If they'd asked more people, they would have had more values in the plot. Then the line would maybe fit better. 'Cause the more data you have, the more you have to base your representation on. And it would be closer to the truth, like, the more people you ask, the more basis you have for saying how it really is.

We see this phase as the students starting to critically assess the likelihood of their own inferences (cf. Makar & Rubin, 2009). When evaluating their initial hypotheses, the students all gave more details on what evidence they based their inferences on, and included more evidence such as trends and variability in the scatterplot, signal (the line through the scatterplot) and noise (distance from the line to observations, both individual distances and overall distance), uncertainty levels in different samples, representativeness, how an outlier would greatly affect small samples like the one in their problem, and alternative explanations to the relationship between the two variables (like studying). Moreover, when the students introduced sample size and variability to their reasoning, it seemed to catalyze their critical thinking on the other aspects mentioned above. The idea that the sample, with its variability and limited size, gives no complete conclusion – *the* key idea in inferential reasoning (Sotos et al., 2007) – seemed to have triggered the students' probabilistic reasoning on alternative explanations, suggestions for improving the model's generalizability, and discussing the relationship between statistical models and reality.

IIR may show problematic for students if they are completely unconnected with statistical properties (Pfannkuch, 2006). However, we here see that the inclusion of statistical ideas contributed to a fruitful development in the students' IIR. In the process of critically reviewing their initial hypotheses, their argumentation gradually took a more probabilistic approach, both through explicitly discussing elements that increase uncertainty and in their wordings in general. The students now to a greater extent included probabilistic wordings like "*it seems like ...*", "*it could be ...*", "*maybe ...*", *"from this plot ..."* and addressed the general uncertainties in drawing statistical inferences.

4.3 Stating a reformulated hypothesis

Following the critical discussion of their own inferences, which we have outlined above, the students all reformulated their initial hypotheses. Though in individual phrasings, their reformulated hypotheses reflected a development in evidence used, degree of generalization and in terms of probabilistic reasoning. Each student's reformulated hypothesis is presented in Table 2.

***Table 2**. The five students' initial and reformulated hypotheses.*

Student	**Initial hypothesis**	**Reformulated hypothesis**
Emma	*One can assume that candy improves how one performs on the exam.*	*If you work hard, then you deserve a treat. And if you work extra hard, you'll get better results.*
Heidi	*The more candy you eat, the better mark you'll get apparently.*	*It seems like there's a trend showing that the more candy you eat, the slightly better exam results you might get.*
Karoline	*It seems like the more candy you eat, the better you do at the exam.*	*If I eat a lot of candy before the exam, then, in theory, I should get better results on the exam.*
Susanne	*It looks like the exam result actually improves, with the more candy you eat. (...) So, maybe it helps to eat candy the week before the exam, then.*	*It might not always be true that eating candy the week before the exam gives a better exam result, but from this plot, it looks like this is the case.*
Linn	*Students who read a lot eat a lot of candy, and they also do well on their exams.*	*There are examples showing that there's not necessarily a connection between candy intake and exam results.*

The increased emphasis on uncertainties within statistical inferences, especially related to representativeness but also to some extent causality, was reflected in the students' reformulated hypotheses. They continued to assume some kind of correlation between candy intake and exam results within the data but were now expressing doubt in whether this applied to the real world (beyond the data). Critically reflecting on the validity and adaption of a model is crucial in linear regression (Garfield & Ben-Zvi, 2008). The students' reformulated hypotheses to a greater extent show awareness of the risk of over-generalizing when drawing inferences that go beyond the data, through formulations like "*a trend*", "*in theory*" and *"from this plot"*, illustrating a more probabilistic approach. Emma and Linn's reformulated hypotheses (and to some extent Susanne) also reveal their concerns regarding claiming causality between the two variables. Through critically reflecting on their own starting point – their initial hypothesis – the students' reasoning developed towards expressing a probabilistic generalization beyond data, which is in line with the key elements of IIR presented by Makar and Rubin (2009).

5 Concluding Discussion

Entering this study, we were curious on how undergraduate students' IIR could develop in an inquiry problem-solving session. In three stages after receiving the problem, we've described and exemplified the development of five students' IIR from a micro-perspective. The creation of an initial hypothesis – although unjustified and deterministic – seemed important for the students' IIR, as it served as a starting point for them to verbally express and further critically evaluate and challenge their inferences. The process of creating initial, tentative hypotheses might in this way have prepared their minds for what was to come and thus made a useful starting point for their IIR. Makar and Rubin (2009) consider the process of generating initial hypotheses as one type of informal inference, and this creative but maybe uncritical first attempt to interpret statistical data has been acknowledged as fruitful for students' reasoning in previous research (e.g Batanero et al., 1997; Ben-Zvi, 2004; Ben-Zvi & Arcavi, 2001). An important aspect of statistical modelling is finding an overall trend to "let students see the signal in the noise" (Bakker & Gravemeijer, 2004, p. 165). We argue that critical reasoning on samples also can facilitate the students in seeing *the noise around the signal*, and that awareness of the noise – variability, shape, and outliers – is crucial in facilitating probabilistic reasoning. Focusing too much on the signal can contribute to overlooking the uncertainties within a statistical model, as illustrated through the students' rather non-probabilistic and un-justified initial hypotheses about the general correlation (and, for four of the five, a claimed causal relationship) between candy intake and exam results. Reflecting on the noise led the students toward recognizing that a sample only provides partial information about a population (cf. Sotos et al., 2007). Although informal and hence not bound to any formal statistical methods, the students' IIR integrated several statistical concepts and ideas, like representativeness, variability, context, correlation, (non-)causality, signals and noise, generalizations, and uncertainty, all in which the students critically and flexibly maneuvered, interrelated, and discussed. Further, they included more evidence and justifications to their arguments and questioned the certainty of their previous inferences. Arguing, justifying, and evaluating one owns thinking anchored in statistical concepts and ideas as shown here, is well-known to foster in-depth learning (Ben-Zvi & Garfield, 2004; Makar & Rubin, 2009, 2018; Zieffler et al., 2008), and in the process of making logical hypotheses or conclusions, one needs to provide persuasive arguments based on data (Ben-Zvi, 2006; Makar & Rubin, 2009).

Most of the previous research on IIR has focused on how it can be a first step towards making *formal* inferences. Our research adds to a recent change in perspective (see Makar & Rubin, 2018) towards valuing the power of informal reasoning without it being a prelude to something more formal, and exploring its possible nature. This study shows one example of the developments that might form in students' informal reasoning in a micro-perspective, as they grapple with making and evaluating their own hypotheses and accompanied arguments in an effort to generalize based on a small sample. IIR can be seen as a relatively new area of interest in statistics research (Makar & Rubin, 2018), meaning that further studies that explore IIR through different perspectives and with different aims, are needed. Our research must be seen as just one piece in a larger puzzle, and more research, including larger studies, is needed to gain thorough understanding of the elements and developments that might occur in students' IIR when grappling with scatterplots.

Further, our research can inform teachers and researchers. We found that enabling and encouraging the students to grapple with *their own* ideas and hypotheses seemed central in the development of their IIR. Based on the developments we saw in the five students' IIR in terms of correlation, we advocate for learning activities that give the students both time and intellectual space to deeply engage in, critically evaluate, and refine, hypotheses made from their own interpretations of what a scatter plot can tell them. This adds to research supporting teaching and learning approaches that encourages critical thinking, inquiry, and inference (e.g. Artigue & Blomhøj, 2013; Ben-Zvi, 2006; Dorier & Maaß, 2020; Garfield & Ben-Zvi, 2008; Konold & Pollatsek, 2002; Makar & Rubin, 2009, 2018; Zieffler et al., 2008). Continuing our call for research that explores students' IIR through different lenses, we aknowledge that the design and orcestration of the activity in our research was motivated by our research aim and methods (think-aloud protocols) and must only be seen as one approach to IIR. For example, it takes place in a artificial context outside the classroom, and the instructors did not take on a teacher role. We welcome research that explore the design and use of learning activities that draw on our suggestions above in a real classroom context. Furthermore, while our study implies that IIR is an approach that promotes critical thinking as it enables and encourages students to express, elaborate on, end evaluate, their own ideas, and thus makes the students accountable for justifying and modifying their hypotheses, the role of the teacher should not be neglected. An idea for future research might be to look into the teacher role, e.g., questioning techniques and orchestration of learning activities that allows the students time and intellectual space but at the same time provokes critical thinking and evaluation of the students' IIR. We acknowledge the micro-perspective of this study. Our results and discussion should be interpreted as suggestions based on the researchers' interpretations of what the five students were sharing in their reasoning. However, we do believe that even through this small sequence, our study gives insight to how grappling with, and combining, statistical knowledge and informal meaning-making, can be catalytic on the development of key elements in IIR. Further, we highlight the importance of inquiry and concurrent probing in teaching, as well as providing the students sufficient time and intellectual space to reason and explore their own ideas.

References

Afflerbach, P., & Johnston, P. (1984). On the use of verbal reports in reading research. *Journal of Reading Behavior, 16*(4), 307-322. https://doi.org/10.1080/10862968409547 52

Artigue, M., & Blomhøj, M. (2013). Conceptualizing inquiry-based education in mathematics. *ZDM Mathematics Education, 45*(6), 797-810. doi: https://doi.org/10.1007/s11858-013-0506-6

Bakker, A., & Gravemeijer, K. P. (2004). Learning to reason about distribution. In D. Ben-Zvi and 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., Estepa, A., & Godino, J. D. (1997). Evolution of students' understanding of statistical association in a computer based teaching environment. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics*: *Proceedings of the 1996 IASE Round Table Conference* (pp. 191-205). International Statistical Institute.

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematics Education in Science and Technology, 25*(4), 527-547. https://doi.org/10.1080/0020739940250406

Ben-Zvi, D. (2004). Reasoning about data analysis. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 121-145). Kluwer. https://doi.org/10.1007/1-4020-2278-6_6

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics* (pp. 1-6). International Statistical Institute.

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. https://doi.org/10.1023/A:1013809201228

Ben-Zvi, D., & Garfield, J. (2004). Statistical Literacy, reasoning and thinking: Goals, definitions and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 3-15). Kluwer. https://doi.org/10.1007/1-4020-2278-6_1

Ben-Zvi, D., Gil, E., & Apel, N. (2007). What is hidden beyond the data? Young students reason and argue about some wider universe. In D. Pratt & J. Ainley (Eds.), *Reasoning about Informal Inferential Statistical Reasoning: A Collection of Current Research Studies. Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy*. University of Warwick.

Bryman, A. (2016). *Social Research Methods (5th ed).* Oxford University Press.

Bråtalien, M., & Naalsund, M. (2021). Undergraduate statistics students' reasoning on simple linear regression. In G. A. Nortvedt, N. F. Buchholtz, J. Fauskanger, F. Hreinsdóttir, M. Hähkiöniemi, B. E. Jessen, J. Kurvits, Y. Liljekvist, M. Misfeldt, M. Naalsund, H. K. Nilsen, G. Pálsdóttir, P. Portaankorva-Koivisto, J. Radišić, & A. Wernberg (Eds.), *Bringing Nordic Mathematics Education into the Future, Preceedings of Norma 20: The Ninth Nordic Conference on Mathematics Education.* (Vol. 13, pp. 17-24). Swedish Society for Research in Mathematics Education.

Dorier, J.-L., & Maaß, K. (2020). Inquiry-based mathematics education. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education*. Springer. https://doi.org/10.1007/978-94-007-4978-8_176

Garfield, J., & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice.* Springer. https://doi.org/10.1007/978-1-4020-8383-9

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259-289. https://doi.org/10.5951/jresematheduc.33.4.0259

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 151-168). International Statistical Institute.

Lapadat, J. C. (2000). Problematizing transcription: Purpose, paradigm and quality. *International Journal of Social Research Methodology, 3*(3), 203-219. https://doi.org/10.1080/13645570050083698

Lithner, J. (2008). A research framework for creative and imitative reasoning. *Educational Studies in Mathematics. 67*(3), 255-276. https://doi.org/10.1007/s10649-007-9104-2

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

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

Petersson, J. (2022). Students' responses to the question: how does a computer do curve fitting? *International Journal of Mathematical Education in Science and Technology*, 1-17. https://doi.org/10.1080/0020739X.2022.2053216

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics. In Salvador, Brazil.* International Statistical Institute.

Powell, A. B., Francisco, J. M., & Maher, C. A. (2003). An analytical model for studying the development of learners' mathematical ideas and reasoning using videotape data. *The Journal of Mathematical Behavior, 22*(4), 405-435. https://doi.org/10.1016/j.jmathb.2003.09.002

Pratt, D., & Ainley, J. (2008). Introducing the special issue on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 3-4. https://doi.org/10.52041/serj.v7i2.466

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98-113. https://doi.org/10.1016/j.edurev.2007.04.001

Stake, R. E. (2003). *Case Studies (2nd ed.).* Sage.

Van Someren, M., Barnard, Y., & Sandberg, J. (1994). *The Think Aloud Method: A Practical Approach to Modelling Cognitive*. Academic Press.

Watson, J. M. (2004). Developing reasoning about samples. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 277-294). Kluwer. https://doi.org/10.1007/1-4020-2278-6_12

Young, K. A. (2005). Direct from the source: the value of 'think-aloud' data in understanding learning. *Journal of Educational Enquiry, 6*(1), 19-32.

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