Student Conceptualizations of Confidence Interval Attributes

Susan Lloyd
The Pennsylvania State University

Neil Hatfield
The Pennsylvania State University

Matthew Beckman
The Pennsylvania State University

Nicole Lazar
The Pennsylvania State University

Interval estimation has become one of the preferred means for communicating the statistical findings of a study. Students, and even experienced professionals, often misinterpret frequentist confidence intervals. As part of the broader study in which this work is situated, we developed an instrument to assess undergraduate students' statistical literacy surrounding confidence intervals. We conducted think-aloud interviews with six students who had recently completed an introductory statistics course. In this work, we qualitatively capture the meanings that students construct about attributes of confidence intervals to inform the results from the instrument. We use grounded theory to model how students conceptualize attributes of confidence intervals and propose evidence of interactions between interrelated ways of thinking. For this paper, we focus on one student's conceptualizations about the confidence level and error attributes of confidence intervals to illustrate our modeling process in detail.

Keywords: confidence intervals, grounded theory, interaction effects, statistics education

Introduction

Hypothesis testing is often presented as a rote multi-step procedure (Horton, 2015; Lock et al., 2017). This training has led many research communities to falsely believe that only statistically significant results are meaningful. The standards adopted by entire disciplines that require statistically significant results for publication perpetuate this belief. As a result, researchers strive to achieve statistical significance, which can lead, and has led, to research malpractice (Collins, 1984; Wasserstein et al., 2019).

Moreover, the outcomes from hypothesis testing are frequently misinterpreted by students and professionals (Hoekstra et al., 2014). Interval estimation has been proposed as a more acceptable alternative since confidence intervals are accessible, readily display the precision of the interval construction method, and are not restricted to a specific null hypothesis (Hoekstra et al., 2014, 2018; Kaplan & Roland, 2022a). However, individuals treat confidence intervals like significance testing when they focus on whether the null value falls in the interval (Wasserstein et al., 2019). As such, it is insufficient for just the *presentation* of the findings to shift from a p-value to a confidence interval; the *interpretation* of the findings must shift as well from that of desiring statistical significance to that of desiring transparency of results. More broadly, there must be a shift from procedural treatment of statistical tools towards building and working with deep, productive meanings and exhibiting quantitative reasoning (Hatfield et al., 2022). Thus, studying individuals' understanding of confidence intervals is imperative.

Literature Review

Individuals, even experienced researchers, grossly misinterpret confidence intervals (Hatfield, 2024; Hoekstra et al., 2014, 2018; Roland & Kaplan, 2022; Saldanha et al., 2024). In fact, even statisticians and authors of introductory statistics textbooks disagree regarding how to interpret a frequentist confidence interval (Hoekstra et al., 2018; Kaplan & Roland, 2022a). These misinterpretations are worsened by the ambiguity surrounding the word ‘confidence’ (Kaplan & Roland, 2022a, 2022b). Morey et al. expand upon these misinterpretations of confidence intervals by identifying the most common confidence fallacies: fundamental confidence fallacy, precision fallacy, likelihood fallacy, and fallacy of acceptance. They describe these misinterpretations as fallacies since they are in opposition to Neyman’s theory of confidence intervals (Morey et al., 2016). The fundamental confidence fallacy deals with making probability claims about the resulting confidence interval. These claims reference the probability or confidence that any realized interval contains the parameter (Hoekstra et al., 2018; Saldanha et al., 2024). These claims place the confidence in one particular interval, rather than the method where it rightfully belongs (Hoekstra et al., 2014, 2018; Saldanha et al., 2024). The precision fallacy deals with belief about precision based on the width of the confidence interval (Hoekstra et al., 2018; Morey et al., 2016). The precision fallacy manifests itself in the belief that a narrow interval gives more knowledge or suggests that the statistic is a better estimate of the parameter.

The likelihood fallacy is thinking that the confidence interval presents a set of reasonable or plausible values for the parameter. If a value falls inside the interval, it is reasonable, otherwise it is unreasonable (Hoekstra et al., 2018; Morey et al., 2016). Thoughts that align with the likelihood fallacy include assigning probability to individual values or thinking that values in the interval are more likely values for the parameter (Saldanha et al., 2024). When interpreting a confidence interval, individuals naturally want to accept a set of values, which is known as the fallacy of acceptance (Hoekstra et al., 2018; Mayo & Spanos, 2006; Morey et al., 2016). Accepting a value as the true parameter because it falls in the interval is an indicator of the fallacy of acceptance. Other problematic meanings of confidence intervals include believing the interval is a descriptive statistic (Hoekstra et al., 2014), is a set of reasonable values for individual cases (Hatfield, 2024) or for the sample statistic (Kaplan & Roland, 2022b).

After discussing some problematic meanings, we now present a productive meaning of confidence intervals. A person with such meaning would view the process of constructing a confidence interval as entailing randomness, leading to variation in the resulting intervals when the process is repeated. They would also have the ability to conceptualize and articulate the distinction between the procedures for, and results from, obtaining a sample and constructing a confidence interval (Hatfield, 2024). Therefore, in a productive interpretation, the confidence is in the method, not in a particular interval, and there is recognition of uncertainty involved in the process (Saldanha et al., 2024).

There are a few assessments in the statistics education literature to measure understanding of confidence intervals (delMas et al., 2006, 2007). For example, the Comprehensive Assessment of Outcomes in Statistics (CAOS) measures statistical literacy, reasoning and thinking on major topics covered in the introductory statistics curriculum, including confidence intervals (delMas et al., 2007). Conventional practices in assessment development resemble a one-factor-at-a-time design in which each item assesses one idea at a time, independent of the other ideas (see delMas et al., 2007; Sabbag et al., 2015; Ziegler, 2014). From a cognitive perspective, ideas do not necessarily develop in isolation of each other. Rather individuals often coordinate multiple ideas; thus, evaluating the interactions between multiple ideas is crucial (Thompson et al., 2014).

Methods

The work we present here is part of a larger project which investigates students’ statistical literacy surrounding confidence intervals after an introductory course. In an on-going feasibility study, we are developing, validating, and analyzing an instrument to evaluate interactions among attributes of confidence intervals. This paper intends to provide qualitative insight into the instrument by addressing the following research question in support of these broader aims: How do undergraduate students conceptualize, and coordinate their meanings for, confidence level and error in the context of confidence intervals?

Drawing upon the literature (e.g., Hatfield, 2024; Hoekstra et al., 2018; Lock et al., 2017) and our own expertise, we identified several attributes of confidence intervals that are often taught in an introductory course: confidence level, error, sample size, p-value, and method. For this paper, we focus on the first two attributes. We operationalize confidence level as the proportion of times the process of constructing a confidence interval results in an interval that captures the parameter of interest (i.e., success rate) and error as a measure of variability associated with the data and with the sample statistics.

Theoretical Perspective

In this study, we adopt an interpretivist epistemology, in particular, radical constructivism (Stinson, 2020), which we use to guide the formation of models of student conceptualizations about attributes of confidence intervals. We consider knowledge to be constructed by students based on their interactions and experiences, both in life and in the classroom (Stinson, 2020; Thompson, 2008; von Glasersfeld, 1995). A radical constructivist approach acknowledges that students construct different meanings or ways of thinking (Goldin, 2000), reflecting student autonomy.

Interview Protocol

Cognitive interviews are used for a multitude of purposes in mathematics and statistics education (Clement, 2000; Goldin, 2000; Theobold et al., 2023). One benefit is that they are useful for observing students' mathematical/statistical thinking processes and approaches (Case & Jacobbe, 2018; Clement, 2000; Goldin, 2000). In this study, we conducted think-aloud interviews with six students who recently completed an introductory statistics course. We interviewed these students separately, on Zoom, for one hour. We asked students to read and discuss their thinking aloud as they interacted with our items. As needed, we asked follow-up questions. We audio recorded the interviews to obtain detailed transcripts.

Data Analysis

To model how students thought about confidence intervals, we used grounded theory. Grounded theory is a qualitative method for deriving a theory to describe a phenomenon of interest in a manner that reasonably stems from the observed data and has a clearly outlined scope of generalizability (Strauss & Corbin, 1990). We used a generative approach in that we developed our codes/theories based on students' demonstrated thinking about confidence level and error rather than having a priori hypotheses about the meanings students have for these attributes (Clement, 2000). To start, we read the transcripts and flagged sections (sentences, paragraphs) corresponding to each of the five attributes. Starting with the confidence level attribute, we identified codes (concepts) for each section to represent the thinking that students conveyed. We drew upon existing literature and our understanding of students' statements as a basis for our theoretical sensitivity and to inform the open coding process (Strauss & Corbin, 1990). Examples of our codes include: confidence level refers to confidence in the method, fundamental confidence fallacy, confidence level is a success/capture rate, articulates the normatively accepted relationship between confidence level and interval width. We grouped (categorized) the codes for each student to create higher order descriptions, encompassing multiple concepts (Strauss & Corbin, 1990). We developed a list of defining properties for each category by iteratively revisiting the original sections and making note of instances which supported (or not) the proposed relationship. Elaborating on the nature of this variation enabled us to construct a more nuanced theory describing the phenomenon of interest (Strauss & Corbin, 1990). For this paper, we focus on just one student (whom we call S3) to illustrate in depth how we model students' conceptualizations about the confidence level and error attributes.

Results

Consider the following excerpt from S3 which is part of an exchange on why they chose a provided option that indicated the width of a 90% confidence interval is narrower than that of a 95% confidence interval (for the same sample) because it has a smaller multiplier (critical z-score), leading to a smaller margin of error. This particular excerpt demonstrates S3's meaning for confidence level: "I'm pretty sure we always just kind of grab multipliers off a table so…I don't really remember…I don't know a whole lot about actually calculating it…based on like confidence." From this excerpt, we observe that S3 has a meaning consistent with a procedural acquisition of the multiplier. This excerpt, along with others coded similarly, lead us to believe that S3 has a meaning consistent with a formulaic treatment of confidence intervals. By this we mean that an individual has an understanding of confidence intervals as a formulaic calculation on individual components. Although this meaning allows S3 to articulate the normatively accepted relationship between confidence level and multiplier (as confidence level increases, the multiplier increases in magnitude), this meaning makes it difficult for S3 to determine how intervals with the *same* confidence level, but constructed using different methods, would differ. Moreover, the following excerpt is in response to an item that asks students to explain, if 100 students replicate a given study, how many of the 100 resulting 90% confidence intervals are expected to capture the parameter of interest. This excerpt reveals that S3 has a meaning for confidence level that is consistent with how we operationalized it:

> Alright so the method in this question method in this case is one we're 90% confident in so…therefore, we would *expect*…for every 100 intervals we would. Well, we would simply multiply the number of intervals by the...rate of correct of a…rate of expected correct, of a[n] expected capture…which would be 90.

Here S3 appears to imagine carrying out a process to produce multiple intervals. Further, S3 connects the confidence level to the notion of a success rate (how many intervals capture the true value). Taking the two instances together, S3 appears to coordinate a repeatable calculational process and anticipated success of the results. This thinking supports S3 in articulating the normatively accepted relationship between confidence level and interval width (as confidence level increases, interval width increases) and meanings for confidence.

Consider the following exchange between S3 and the interviewer on a multiple-choice item that demonstrates S3's meaning for error:

> *S3:* Using the empirical rule, the 95% confidence interval for the mean difference in plant heights would take the following form [statistic $\pm$ measure of error expression]. Okay, the empirical rule...in other words 68, 95, 99 point…something.
>
> *Interviewer:* Mhm 7.

S3: Yeah, yeah one standard error...captures about 68%, 2 standard errors captures 95%. So multiplier is 2 should be 2 in this case. So that…(5 second pause)…okay, margin of error. Right. You said margin of error is the…the radius right? That's what I said. So.

Interviewer: Everything to the right of the plus or minus sign.

S3: Yeah, yeah, exactly. So that's the 2…so…so in other words, that would be…that's the result of multiplying the multiplier and the standard error.

From this exchange, we observe that S3 conveys the margin of error as being a distance for the interval, specifically, a radius, and getting the error via a product of components. There is still a formulaic aspect to S3's meaning for error. S3 routinely comes back to the multiplication aspect as well as the Empirical Rule. While S3 can articulate error as a distance quantity, they tend to focus more on the calculations of and with error than articulating any underlying quantitative relationships involving error. Even so, S3 is able to articulate the normatively accepted relationship between standard error and interval width (as standard error decreases, interval width decreases) but lacks sureness in their meaning of error terms apart from the formulas. We note that S3 conveys that the standard error is a property of the data instead of the statistic, and they imagine deviation as referring to differences between the intervals.

When discussing both confidence level *and* error, S3 appears to coordinate their meanings. Consider the following excerpt from S3 which details their reasoning on a multiple-choice item assessing how the standard error and margin of error for a 90% and 95% bootstrap confidence interval (constructed using the same bootstrapped sampling distribution) would compare:

> So standard error is that a property of the? That's I kinda wanna say standard error is a property of the data itself? So I don't think that would differ between the two methods [method of constructing a 90%, 95% confidence interval] but the margin of error *should* differ because that's a matter of how much leeway we give ourselves, how much confidence we're willing to sacrifice.

S3 appears to use their meaning of confidence level as the method's success rate, and their meaning of standard error as a property of the data, to construct a meaning for error that no longer has the distance aspect previously mentioned. Rather, this meaning for error is in-line with the colloquial use of the word error, as in the interval has the property of making mistakes.

Discussion

A consistent meaning that students appear to have is a formulaic treatment of confidence intervals. It would be of interest to study whether a formulaic approach is an efficient way to solve the respective items and/or whether this meaning is influenced by the presentation of the general formula for a confidence interval throughout the instrument. Moreover, S3's conceptualizations of confidence level and error display the interactions between their ways of thinking about these attributes of confidence intervals. This preliminary evidence of interactions among attributes of confidence intervals justifies the need to efficiently estimate these interaction effects and indicates instances of quantitative reasoning (Hatfield et al., 2022) and potentially covariational reasoning (Carlson et al., 2002; Thompson & Carlson, 2017; Zieffler & Garfield, 2009). There is further insight to be gleaned on this front as we have five other students for whom we can model their conceptualizations of the confidence level and error attributes. We also have three additional attributes of confidence intervals (sample size, p-value, method) to analyze. By understanding how students coordinate meanings for attributes related to confidence intervals, instructors can achieve greater clarity on what changes are needed in curriculum or assessment to foster statistical literacy on the topic of confidence intervals.

Acknowledgments

S. Lloyd was supported by the Eberly College of Science Janet L. Norwood Science Achievement Graduate Scholarship in Statistics at the Pennsylvania State University.

References

Carlson, M., Jacobs, S., Coe, E., Larsen, S., & Hsu, E. (2002). Applying covariational reasoning while modeling dynamic events: A framework and a study. *Journal for Research in Mathematics Education, 33*(5), 352–378. https://doi.org/10.2307/4149958

Case, C., & Jacobbe, T. (2018). A framework to characterize student difficulties in learning inference from a simulation-based approach. *Statistics Education Research Journal*, *17*(2), 9–29. https://doi.org/10.52041/serj.v17i2.156

Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 547–589). Lawrence Erlbaum Associates, Inc.

Collins, R. (1984). Statistics versus words. *Sociological Theory, 2*, 329–362. https://doi.org/10.2307/223353

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28–58. http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf

delMas, R., Ooms, A., Garfield, J., & Chance, B. (2006). Assessing students' statistical reasoning. *Proceedings of the 7th International Conference on Teaching Statistics (ICOTS7, 2006), Salvador, Brazil.* ISI/IASE. https://iase-web.org/documents/papers/icots7/6D3_DELM.pdf?1402524965

Goldin, G. A. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 517–545). Lawrence Erlbaum Associates, Inc.

Hatfield, N. (2024). Using constructivism to develop an agenda of research in stochastics education research. In P. C Dawkins, A. J. Hackenberg, & A. Norton (Eds.), *Piaget's genetic epistemology for mathematics education research* (pp. 605–611). Springer Nature Switzerland AG.

Hatfield, N. J., Saldanha, L., Primi, C., & Chernoff, E. (2022). Quantitative reasoning and conceptual analysis as a framework for teaching and learning probability. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering & educating today's learners in statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11, 2022), Rosario, Argentina.* ISI/IASE. https://iase-web.org/icots/11/proceedings/pdfs/ICOTS11_236_HATFIELD.pdf

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Hoekstra, R., Morey, R. D., & Wagenmakers, E.-J. (2018). Improving the interpretation of confidence and credible intervals. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_8A2.pdf?1531364291

Horton, N. J. (2015). Challenges and opportunities for statistics and statistical education: Looking back, looking forward. *The American Statistician, 69*(2), 138–145. https://doi.org/10.1080/00031305.2015.1032435

Kaplan, J. J., & Roland, K. E. (2022). Confidence means what?!?: Lexical ambiguity in the interpretation of confidence intervals. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering & educating today's learners in statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11, 2022), Rosario, Argentina.* ISI/IASE. https://iase-web.org/icots/11/proceedings/pdfs/ICOTS11_268_KAPLAN.pdf?1669865547

Kaplan, J. J., & Roland, K. E. (2022). Content and pedagogical knowledge for teaching confidence intervals in a post p < 0.05 world. In P. C. Short, H. Henson, & J. R. McConnell (Eds.), *Age of inference: Cultivating a scientific mindset* (pp. 317–346). Information Age Publishing.

Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2017). *Statistics: Unlocking the power of data* (2nd ed.). John Wiley & Sons.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *The British Journal for the Philosophy of Science, 57*(2), 323–357. https://www.jstor.org/stable/3873470

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Roland, K. E., & Kaplan, J. J. (2022). Brody and Jamie's colleague: The difference in confidence interval estimators and estimates. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering & educating today's learners in statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11, 2022), Rosario, Argentina.* ISI/IASE. https://iase-web.org/icots/11/proceedings/pdfs/ICOTS11_264_ROLAND.pdf?1669865546

Sabbag, A., Garfield, J., & Zieffler, A. (2015). Quality assessments in statistics education: A focus on the GOALS instrument. In M. A. Sorto (Ed.), *Advances in statistics education: Developments, experiences and assessments. Proceedings of the Satellite Conference of the International Association for Statistical Education (IASE, 2015), Rio de Janeiro, Brazil*. ISI/IASE. http://iase-web.org/documents/papers/sat2015/IASE2015%20Satellite%2041_SABBAG.pdf

Saldanha, L., Hatfield, N. J., Chernoff, E. J., & Primi, C. (2024). *The learning and teaching of statistics and probability: A perspective rooted in quantitative reasoning and conceptual coherence.* Routledge.

Stinson, D. W. (2020). Scholars before researchers: Philosophical considerations in the preparation of mathematics education researchers. *Philosophy of Mathematics Education Journal, 36,* 1–18. https://education.exeter.ac.uk/research/centres/stem/publications/pmej/pome36/index.html

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Sage Publications.

Theobold, A. S., Wickstrom, M. H., & Hancock, S. A. (2023). Coding code: Qualitative methods for investigating data science skills. *Journal of Statistics and Data Science Education, 32*(2), 161–173. https://doi.org/10.1080/26939169.2023.2277847

Thompson, P. W. (2008). Conceptual analysis of mathematical ideas: Some spadework at the foundation of mathematics education. In O. Figueras, J. L. Cortina, S. Alatorre, T. Rojano, & A. Sépulveda (Eds.), *Plenary paper presented at the Annual Meeting of the International Group for the Psychology of Mathematics Education,* (Vol. 1, pp. 31–49). PME.

Thompson, P. W., & Carlson, M. P. (2017). Variation, covariation, and functions: Foundational ways of thinking mathematically. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 421–456). National Council of Teachers of Mathematics.

Thompson, P. W., Carlson, M. P., Byerley, C., & Hatfield, N. (2014). Schemes for thinking with magnitudes: A hypothesis about foundational reasoning abilities in algebra. In K. C. Moore, L. P. Steffe & L. L. Hatfield (Eds.), *Epistemic algebra students: Emerging models of students' algebraic knowing*. WISDOMe Monographs (Vol. 4, pp. 1–24). Laramie, WY: University of Wyoming.

von Glasersfeld, E. (1995). *Radical constructivism: A way of knowing and learning.* Falmer Press.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$." *The American Statistician, 73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Zieffler, A. S., & Garfield, J. B. (2009). Modeling the growth of students' covariational reasoning during an introductory statistics course. *Statistics Education Research Journal, 8*(1), 7–31. https://doi.org/10.52041/serj.v8i1.455

Ziegler, L. (2014). Reconceptualizing statistical literacy: Developing an assessment for the modern introductory statistics course. Retrieved from the University Digital Conservancy, https://hdl.handle.net/11299/165153