STUDENTS' ARTICULATION OF UNCERTAINTY ABOUT BIG DATA IN IMA LEARNING ENVIRONMENT

Ronit Gafny and Dani Ben-Zvi
University of Haifa, Israel
ronit.gafny@edtech.haifa.ac.il

This study examines students' expressions of uncertainty while interacting with classic and nontraditional big data analyses. The study was designed according to the integrated modeling approach (IMA), which was found to be suitable for the development of reasoning with uncertainty in a classic data setting. Over the course of the activity, 87 expressions of uncertainty were identified. A total of ten types of uncertainty expressions were identified: eight occurred during big data activities, and five occurred during classic data activities. Furthermore, a conceptual framework for describing novices' reasoning with uncertainty with big data has been developed. The study also illustrates the pedagogical potential of implementing IMA in big data settings and combining classic data with big data investigations.

PURPOSE

The motivation for this research is the recognition of the importance and impact of today's big data era, especially on the lives of young people born into the digital age. Thus, developing reasoning with big data is essential for all populations (young students, older students, teachers, etc.) to better understand and manage their world. The main task of statistics education is to provide students with conceptual frameworks and practical skills to better serve them in their future world (Wild et al., 2018). This study aims to extend existing research insights in statistics education regarding informal statistical reasoning to the big data and data science worlds. One main challenge is preparing students to attend to the unique uncertainties embedded in big data. A deeper understanding of students' reasoning with various uncertainties as they engage with different data types can provide pedagogical insights into better supporting them.

The research questions. In the context of a learning sequence that combines a classic data exploration with big data exploration, the main research question is: *What are the characteristics of graduate students' reasoning with uncertainty in an integrated modeling approach (IMA) designed environment?* The sub-questions are: (a) What articulations of uncertainty can graduate students express? and (b) What characterizes these expressions?

THEORETICAL BACKGROUND

Big data. In today's world, high connectivity produces a huge amount of data, often from sensors, smartphones, digital apps, social media, citizen science projects, and many other unconventional sources. This data can be incomplete, imprecise, misleading, or even inaccurate. These new types of data are in sharp contrast to classic data used in school curricula, which are mostly well-documented, precise, complete, and, in most cases, sampled according to a methodological sampling scheme. There are various definitions for big data. The key characteristics that are typically suggested: Volume, Velocity, and Variety, Value, Veracity, Variability and Complexity, and Value. Some of the V's suggest different types of uncertainties associated with big data.

Uncertainty. Uncertainty is a key aspect of statistics, typically accounted for by the use of probability. Uncertainty is no less central in the world of big data. Uncertainty is embedded throughout the entire data investigation process (Hariri et al., 2019). Statistical uncertainty is related to "data" and "chance," both phenomena treated by statistics and probability, respectively (Moore, 1990). Statistics tends to focus on randomness-related uncertainty, whereas probability allows measurement of the level of uncertainty that characterizes a phenomenon. Classic data analysis classifies expressions of uncertainty into two main categories. The first, *statistical uncertainty*, results from the fact that inferences based on a random sample for describing a larger population require mediation between two seemingly opposing ideas. One is that a sample can represent a population, and the other is that samples of similar size can display different images regarding the same phenomenon, which is also known as sampling variation (Manor & Ben-Zvi, 2015). The second major type, *contextual uncertainty*, results from conflicts between students' contextual knowledge, specifically regarding what they reflect about

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.* International Association for Statistical Education. iase-web.org

the investigated phenomenon, and what the data tell them (Manor et al., 2013). Contextual uncertainty can arise from partial or missing information about how an investigated phenomenon behaves.

The customary classification in data science is different. In relation to models, it distinguishes between aleatory and epistemic uncertainties. Generally, aleatory uncertainty refers to the notion of randomness, i.e., the variability in data that results from the randomness effect. Controversy, a decision-maker's ignorance of the situation, results in epistemic uncertainty (Hüllermeier & Waegeman, 2021). The two classifications are somewhat similar, but they focus on different aspects. The concept of *statistical uncertainty* is usually associated with sampling variability (Manor & Ben-Zvi, 2015), at least in classic data explorations in schools' curricula. Although aleatory uncertainty is broader than sample variation, it also refers to any variation that arises from the stochastic behavior of the data, such as natural variability, accidental variability, and measurement variability (Dvir & Ben-Zvi, 2021). *Contextual uncertainty* often focuses on what is known or not known about a phenomenon alone. Conversely, *epistemic uncertainty* may refer to much broader aspects of lack of knowledge, such as knowledge regarding the phenomena, the data, the statistical procedures regarding the data, the tools used to analyze the data, etc.

Integrated Modeling Approach (IMA). The IMA was developed to guide the design and analysis of data investigation tasks and deepen students' reasoning with data and chance, sampling, and modeling (Manor & Ben-Zvi, 2017). The IMA suggests integrating real-world data inquiries with probabilistic modeling activities that allow students to perform in-depth examinations of the uncertainty-related considerations they raise during their real-data investigations. Consequently, it provides fertile ground for the expression of uncertainty (Dvir & Ben-Zvi, 2021).

DESIGN

Participants. This case study is a part of the Connections Project—a longitudinal research project aiming to develop an inquiry-based and technology-enhanced statistics learning environment. The current study focuses on data collected in 2020 as part of the "Developing Statistical Reasoning in Learning Communities" graduate course held at the Faculty of Education, University of Haifa. The study examined a summary task given to the course participants and focused on two of the ten summary assignment reports collected.

Method. A qualitative methodology was chosen because the study aimed to provide a detailed description of an unknown phenomenon—specifically, a case study approach. The data analysis was conducted according to the interpretive micro-analysis approach, attending to the genesis of both verbal articulations and gestures as well as their meaning in their context. The validation of the study is conducted through triangulation, which is an examination of interpretations from multiple perspectives.

The Summary Task. The summary task that was designed according to the integrated modeling approach (IMA) consisted of three investigation cycles. The multi-cycle was designed to summon students' multiple experiences with different types of data sets—classic data sets and big data set. The subject of the investigation of all cycles was the Radon gas. The uniform Radon context helps compare expressions of uncertainty and attribution of differences, if found, to the nature of the kind of data studied. The first cycle included a dataset of 29 Radon measurements sampled by students in residential homes in the Haifa area in Israel. The second cycle included a dataset of 450 Radon measurements sampled in homes throughout Israel. The third and final cycle dealt with big data. This data set contained approximately 75,000 Radon samples, which were taken from homes in the state of Minnesota in the United States. The file is relatively "dirty" in that it has deficiencies, duplicates, ambiguities, and different metrics than the Israeli data. It is also rather "heavy" and difficult to handle with the tools students typically work with, such as TinkerPlots (Konold & Miller, 2015) and CODAP (https://codap.concord.org/). Students received the original data file as well as a cleaner version of the data. The original 75,000 cases file was intended to familiarize students with big data and demonstrate its challenges. In this activity, students were asked to examine the original file and answer questions related to it. The cleaner version included a random sample of 5,000 cases drawn from the original big data file. It was designed to assist the students in analyzing the data and modeling it in the same manner and with the same tools that they used to analyze classic data. The cleaning involved only merging categories written differently and deleting irrelevant ones. Without this superficial cleaning and the reduction of the number of cases, TinkerPlots tools would not be usable.

The data analysis cycle consisted of four stages. The initial stage involved familiarizing and reviewing the original big data file and answering questions about its attributes and reliability. The second stage involved examining and exploring the reduced and cleaner data set to make informal inferences about the research question and research conjecture the students had formulated. The design of this task aims to create a link between the second cycle that deals with classic data and the third cycle that deals with big data. The link was established by an opening question in which students were asked to examine whether they could answer their previous research question based on the new data. In the third phase, students were asked to build a probabilistic model of their conjecture using the TinkerPlots Sampler and to compare their new model with the models they had built during the previous cycles. The final stage involved students in reflecting on their big data experience.

The pedagogical purpose of the summary task was to enable students to experiment with reasoning informally with statistical models and modeling and to give them a first-hand experience with the complexity of exploring big data. The research goal accordingly is to examine the expressions of uncertainty that students articulated in the various cycles.

RESULTS

In total, eighty-seven expressions of uncertainty were articulated by the participating students. The study identified ten different types of uncertainty expressions, eight of which were expressed during the big data activities and five during the classic data activities (three overlapped). In addition, the findings indicated that classical analysis had elicited expressions of aleatory uncertainty and epistemic-contextual expressions. The big data activities elicited aleatory uncertainty expressions as well as epistemic-statistical uncertainty. Examination of the types of aleatory uncertainty that appeared in both big data and classical data investigations showed they had similar frequencies in each type of activity and could be classified into the same subtypes: sample size, unsystematic variance, and statistical modeling. This indicates that during initial experiences with big data analysis, aleatory concerns are no less dominant than in classical data analysis. The centrality of sampling-related considerations in classical data investigations explains the prevalence of aleatory expressions in this type of research. With the shift in focus from sampling to a data-driven exploration characterizing big data investigations (Cao, 2018), the importance and centrality of aleatory considerations have not yet been thoroughly explored. The study suggests that despite the shift in focus in big data, aleatory considerations still characterize novices' reasoning with big data.

The classical investigation activities evoked, in addition to aleatory expressions of uncertainty, types of epistemic-contextual uncertainty. This type of uncertainty was unique to the classical investigations and did not arise in big data activities. The absence of contextual expressions of uncertainty in big data investigations is unclear and raises questions about the role of the context in big data investigations for inexperienced students and the role of using contextual knowledge in big data analysis. The big data activities were characterized by, in addition to aleatory types of uncertainty, Epistemic-Statistical uncertainties, which were not expressed during the classical data investigations. These expressions revealed unique uncertainties and considerations that characterized the students' reasoning with big data, such as uncertainties related to the origins of the data, the data quality, non-apriori research question, data load, research procedure complexity, and research tools. The study's findings further point to a specific type of uncertainty that was consistently absent from the students' expressions, the uncertainty that stems from the incomplete nature of big data. This was even though this type of uncertainty is considered central and dominant in the world of big data, and most of the mathematical procedures that characterize the handling of big data are designed to deal with it (Wang & Zhai, 2017).

The study also demonstrates the pedagogical potential of using the IMA approach while adapting it to big data and combining classic data and big data investigations into a single activity sequence. The learning sequence the students engaged with seems to prepare and help them for big data investigations (engaging in aleatory aspects), but also added difficulties in the transition from a classic data investigation paradigm to a paradigm of big data investigations (mostly epistemic-statistical aspects). Illuminating these difficulties can help shape future sequences of activities, which may constructively support students' initial challenges with exploring big data.

CONCLUSION

This study is a pioneering and relatively early attempt to lay the groundwork for a basic framework of categorizing the types of uncertainty that can be ascribed to big data activities compared to classic counterpart activities. In conclusion, this study proposes a conceptual framework to describe novices' reasoning with uncertainty in big data. The framework proposes a new classification of types and sub-types of uncertainties that can constitute a preliminary infrastructure for future research and be used to analyze and characterize students' reasoning with big data. In addition, the study demonstrates the pedagogical potential of using the IMA approach while adapting it to big data and combining classical data and big data investigations into a single activity sequence.

REFERENCES

Cao, L. (2018). *Data science thinking.* Springer International Publishing. https://doi.org/10.1007/978-3-319-95092-1

Dvir, M., & Ben-Zvi, D. (2021). Informal statistical models and modeling. *Mathematical Thinking and Learning*. Advance online publication. https://doi.org/10.1080/10986065.2021.1925842

Hariri, R., Fredericks, H., & Bowers, E. (2019). Uncertainty in big data analytics: Survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), 1–16. https://doi.org/10.1186/s40537-019-0206-3

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Konlod, C., & Miller, C. (2015). *TinkerPlots* (Version 2.3.1) [Computer software]. University of Massachusetts.

Manor, H., & Ben-Zvi, D. (2015). Students' emergent articulations of models and modeling in making informal statistical inferences. In D. Ben-Zvi & K. Makar (Eds.), *Reasoning about models and modelling in the context of informal statistical inferences.* Proceedings of the *Ninth International research rorum on Statistical Reasoning, Thinking, and Literacy (SRTL–*9) (pp. 107–117). University of Paderborn.

Manor, H., & Ben-Zvi, D. (2017). Students' emergent articulations of models and modeling in making informal statistical inferences. *Statistics Education Research Journal*, *16*(2), 116–143. https://doi.org/10.52041/serj.v16i2.187

Manor, H., Ben-Zvi, D., & Aridor, K. (2013). Students' emergent reasoning about uncertainty exploring sampling distributions in an "integrated approach." In J. Garfield (Ed.), P*roceedings of the Eighth International research forum on Statistical Reasoning, Thinking, and Literacy (SRTL8)* (pp. 18–33). University of Minnesota.

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants: A new approach to numeracy* (pp. 95–137). National Academy of Sciences.

Wang, X. Z., & Zhai, J. (2017). *Learning with uncertainty.* CRC Press. https://doi.org/10.1201/9781315370699

Wild, C. J., Utts, J. M., & Horton, N. J. (2018). What is statistics? In D. Ben-Zvi, K. Maker, & J. Garfield (Eds.), *International handbook of research in statistics education.* Springer. https://doi.org/10.1007/978-3-319-66195-7_1