InSTEP

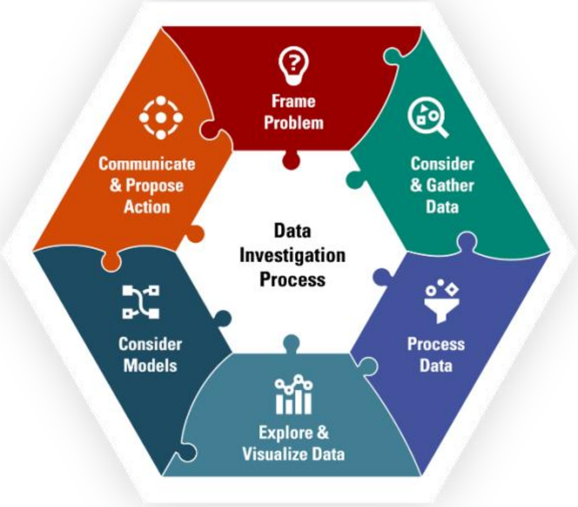
The Data Investigation Process

Today, the ability to make sense of data is essential. K-12 students need educational experiences that can assist them in developing data literacy for global citizenry, and career and college pathways related to statistics and data science (e.g., Engel, 2017; Gould, 2017). Statistics and practices with data are included in standards and goals across the K-12 curriculum. Science puts a heavy emphasis on reasoning from and with data to understand scientific phenomena. English and Social Studies include ensuring data and appropriate visualizations are used to support arguments as well as to understand historical trends, and often inequities, in social structures of our world. Mathematics curricula emphasize measurement data and graphic displays in elementary grades and a strong focus on statistics and probability from grades 6-12. Throughout K-12, students should develop a practice of using data in investigations of real-world phenomena through processes that will prepare them to be data-literate citizens and open doors for data-intensive career pathways in sciences, technology, engineering, journalism, medicine, sports analytics, business, mathematics, statistics, and data science.

For many years, experts in statistics education have described processes used during data-intensive investigations that can guide curriculum and assessment practices in K-12 classrooms. One of the most common is a four-phase investigative cycle, the PCAI model (e.g., Friel et al., 2006; Graham, 1987) or statistical problem-solving process (Franklin et al., 2007; Bargagliotti, et al., 2020), which involves four phases that may or may not be linear: posing a question (P), collecting or considering data (C), analyzing data (A), and interpreting results (I). Another common framework developed by Wild and Pfannkuch (1999) proposes a five-stage investigative cycle: Problem, Plan, Data, Analysis, and Conclusion (PPDAC). Such frameworks often describe other important aspects, such as attention to variability, uncertainty, informal inference, and data as a distribution (Bargagliotti et al., 2020; Franklin et al., 2007; Friel et al., 2006; Lee & Tran, 2015; Wild & Pfannkuch, 1999), highlighting the importance of context. Lee and Tran underscore other statistical habits of mind such as ensuring best measures of an attribute, attending to sampling issues, and using multiple visual and numerical representations to make sense of data. Both Lee and Tran and Wild and Pfannkuch point to the importance of being a skeptic throughout. These are like processes and practices used in science such as asking questions and defining problems, developing and using models, planning and engaging in investigations, analyzing and interpreting data, using computational thinking, arguing from evidence, and constructing explanations (National Research Council, 2012).

With the rise of data science and business analytics in the workplace (Donoho, 2017), several have described the processes used by data scientists and others who work with data to make informed decisions. For example, the career profile for a big-data-enabled specialist describes seven key duties: defines the problem with stakeholders, wrangles data, manages data resources, develops methods and tools, analyzes data, communicates data, and engages in professional development (Education Development Center, 2014). While AJ Goldstein was learning to be a data scientist, he wrote about deconstructing the data science process into 6 essential parts using less technical terms (Goldstein, 2017). With data science education becoming a major world-wide effort, the International Data Science in Schools Project [IDSSP] (2019) released a curriculum framework for guiding K-12 schools in how to introduce students to ways to learn with and from data. They describe a basic cycle of learning from data as including: problem elicitation and formulation, getting the data, exploring data, analyzing data, and communicating results. In addition, dispositions crucial to productively investigating data have been identified (Wild & Pfannkuch, 1999): imagination, curiosity and awareness, openness, engagement, being logical, propensity to seek deeper meaning, and perseverance.

Drawing on the work of those in statistics education and incorporating ideas from data scientists and other professionals, we propose a **Data Investigation Process** that involves six phases. While engaging in the process may be linear, it is often non-linear and dynamic in nature. The diagram illustrates how the six phases fit together like pieces of a puzzle and are all needed for a holistic and productive approach to data investigations. By engaging in and connecting various phases, investigators can make sense of a real-world issue through data and make evidence-based claims and inferences to propose solutions to a problem. Sense-making and interpretation occurs throughout the entire process.



The six phases are briefly described below:

When you ***Frame the Problem***, you consider the context of the real-world phenomena and broader issues that are framing the problem. In considering the variability inherent in the context, you pose one or more investigative questions that could use statistical approaches to answer your question(s). Problems and questions may need to be revised or refocused based on the work in other phases.

During the ***Consider and Gather Data*** phase, you need to consider the types of data that are needed to answer your question, as well as other issues related to collection, design, methods, bias and ethical concerns.

In the ***Process Data*** phase, you consider strategies and techniques for processing and structuring data, such as obtaining data in a usable format and what to do about possible erroneous/invalid and missing data. You also consider processes that may help focus your investigation, such as merging, sorting, grouping, transforming, or filtering your data.

When you ***Explore and Visualize Data***, you create data visualizations (e.g., graphs, images, diagrams), which may be dynamic, and statistical measures that will help you explore and reason about the data in relation to your question and context, as well as look for relationships among variables, patterns and trends.

When you ***Consider Models***, you explore and select the models that help address your problem or answer your questions (e.g., statistical measures, data visualizations (static and dynamic), predictive models, distribution models), considering variability and uncertainty.

In the ***Communicate and Propose Action*** phase, you connect and interpret results and models to the context and make evidence-based claims in relation to a broader problem and specific investigative question. You devise and enact a strategy to convey and interpret your results and models to your audience. You also make recommendations or take action based on the recommendations from your evidence to solve your problem or question, which may include revisiting data from a new perspective or collecting additional data.

It is important to note that the phases of ***Explore and Visualize Data*** and ***Consider Models*** are two shades of the same color in the diagram. This is purposeful to indicate two phases are highly connected and that there is often back and forth and simultaneous analytic considerations when working in these phases--what others have collectively labeled as "analyze data" (e.g., Franklin et al., 2007; Wild & Pfannkuch, 1999). In our framework, we want to draw explicit attention to the role of exploration, visualization, and modeling as key aspects of analyzing data.

While investigations may proceed linearly, all investigations do not emerge and proceed in this way. For example, you may begin with a set of data that has already been collected and do some preliminary exploration and visualization of data, often called exploratory data analysis (EDA). From what is noticed, you may go back to Consider and Gather Data to consider the data source, make sense of different measures, and decide to use different strategies to Process Data in meaningful ways. You may then dive into resources to Frame the Problem by making sense of the bigger context that the data represent and pose a targeted statistical question involving only a few variables in the data set. From there, the appropriate data for the variables of interest would be selected, and you may proceed to Consider Models and require additional work in the Explore and Visualize Data phase. Deciding how to Communicate and Propose Actions may spark new or additional questions to require further investigation with data at hand or require additional data collection and processing.

For more details about the data investigation process, please see the *Thinking Through a Data Investigation* resource. Available at: http://cdn.instepwithdata.org/ThinkingDataInvestigationProcess.pdf

To cite this document:

Hollylynne S. Lee, Gemma F. Mojica, Emily Thrasher, and Zachary Vaskalis. (2020). The data investigation process, In Invigorating Statistics Teacher Education through Professional Online Learning, Friday Institute for Educational Innovation: NC State University. Available at: http://cdn.instepwithdata.org/DataInvestigationProcess.pdf

References

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K-12 Guidelines for assessment and instruction in statistics education (GAISE) report II*. American Statistical Association and National Council of Teachers of Mathematics.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766. https://doi.org/10.1080/10618600.2017.1384734

Education Development Center. (2014). *Big-data-enabled specialists career profile*. http://oceansofdata.org/our-work/profile-big-data-enabled-specialist.

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, *16*(1), 44-49. https://iase-web.org/documents/SERJ/SERJ16(1)_Engel.pdf?1498680968

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report*. American Statistical Association.

Friel, S., O'Connor, W., & Mamer, J. (2006). More than "Meanmedianmode" and a bar graph: What's needed to have a statistical conversation? In G. Burrill and P. Elliott (Eds.), *Thinking and Reasoning with Data and Chance: Sixty-eighth Yearbook* (pp. 117–137). National Council of Teachers of Mathematics.

Goldstein, A. (2017, January 14). *Deconstructing data science: Breaking the complex craft into it's simplest parts*. Mission.org. https://medium.com/the-mission/deconstructing-data-science-breaking-the-complex-craft-into-its-simplest-parts-15b15420df21

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22-25. http://iase-web.org/documents/SERJ/SERJ16(1)_Gould.pdf

Graham, A. T. (1987). *Statistical investigations in the secondary school*. Cambridge University Press.

International Data Science in Schools Project Curriculum Team (2019). *Curriculum frameworks for Introductory Data Science*, http://idssp.org/files/IDSSP_Frameworks_1.0.pdf.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. The National Academies Press. https://doi.org/10.17226/13165

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International statistical review*, *67*(3), 223-248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x