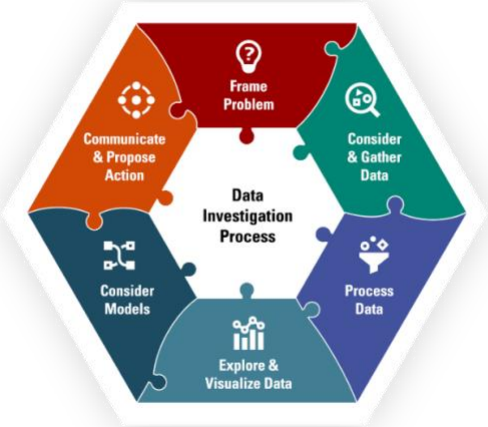
InSTEP

# Thinking Through the Data Investigation Process

Experts in statistics education (e.g., Franklin et al., 2007; Friel et al., 2006; Graham, 1987), as well as data scientists and others who work with data (e.g., Education Development Center, 2014; Goldstein, 2017), have described processes used during data-intensive investigations. We propose a new framework, the **Data Investigation Process**, that brings together fundamental practices and processes from these fields. For a broader description of the data investigation process, see *The Data Investigation Process* resource.




In this document, we identify key considerations to guide thinking and actions for data investigations, where the goal of an investigation is to answer a statistical question within a context to communicate approaches and solutions to a problem based on evidence. This process is composed of six phases: Frame the Problem, Consider and Gather Data, Process Data, Explore & Visualize Data, Consider Models, and Communicate & Propose Action. The jigsaw-like diagram illustrates that each of the phases come together as essential aspects of the Data Investigation Process to explore a real-world phenomenon and make evidence-based claims with data. While it is possible to proceed through an investigation linearly, the process is often non-linear and dynamic in nature, where phases are revisited throughout. To help unpack each of the six phases, the table that begins on p. 2 provides considerations for engaging in data investigations.

While these considerations can be used to guide students' thinking and actions as they engage in an investigation with data, they can also be used as a reflection tool for teachers in planning opportunities to engage students in data-intensive activities. For example, the questions to consider in the table could be used to guide a teacher's planning in designing, identifying or modifying data rich tasks. While it is not intended that students provide a detailed response to every question

when planning or engaging in a data investigation, reflecting on these questions will develop productive habits of mind and data and statistical practices.

Many important aspects of investigating data are highlighted in the table, 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). Context plays an important role in the data investigation process and should be considered throughout. 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 are essential for conducting a productive data investigation process (Lee & Tran, 2015). Both Lee and Tran and Wild and Pfannkuch point to the importance of being a skeptic throughout. It is important to be curious, creative, intuitive, persistent and resilient (Wild & Pfannkuch, 1999), and to communicate and collaborate (IDSSP, 2019); Wild & Pfannkuch, 1999).

To help identify the type of thinking and actions that lead to productive data investigations, we propose considerations for each phase. While shown as two separate phases, Explore and Visualize Data and Consider Models, are highly connected and there are often back and forth and simultaneous analytic considerations when working in these phases--what others have collectively labeled as “analyze data”. At the top of the table, we highlight overarching ideas to be considered throughout a data investigation.

**Considerations for Engaging in a Data Investigation**

**Throughout an investigation, use the following key considerations and dispositions:**

- How are you making sense of the data with respect to the real-world phenomena/context and investigative questions (i.e., engaging in interpretation throughout the process)?
- What is the role of technology? How can it be best used to facilitate your work?
- Are you attending to variability in data and uncertainty in models and claims?
- What biases may be in your data and what biases, experiences, or perspectives do you bring to an investigation that could enhance or negatively impact your work within a context? What privacy or ethical issues need to be considered?
- Do you need to seek out other expertise or find information about the context to inform your work and interpretations?

- Are you being curious, creative and intuitive in your approaches?
- Are you being skeptical as you examine data and the claims and actions that can be proposed?
- How are you communicating and collaborating with team members and/or clients or stakeholders?
- Are you being persistent and resilient in your problem solving when you need to overcome difficult obstacles in an investigation?

**Considerations for Each Phase of the Data Investigation Process**

## Frame the Problem

**Consider the context of the problem.**

- What is the context?
- What is the issue of interest within this context?
- What background information is needed? What resources are available to better understand the context of the problem?
- What is the broader purpose of the investigation? Why is this problem important to consider?
- What kind of data is available in this discipline/context?

**Posing investigative question(s).**

- What statistical questions are you addressing?
- Is the statistical question appropriate for the context of the problem?
- Do these questions anticipate variability?
- Will the question(s) lead to a productive investigation?
- What strategies could potentially be used to answer the question(s)? What types of data are needed to use these strategies?
- What model assumptions should be considered about underlying populations or processes related to the context?

## Consider and Gather Data

**Consider types of data available and collection.**

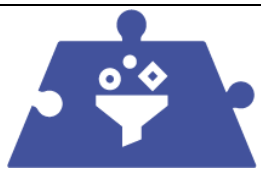
- If data has been collected and readily available:
  - What is the data source(s)?
  - How is the data stored and accessed?
  - What observational or experimental designs were used to collect this data?
  - What are the variables? Which variables are relevant to addressing the issue or answer the question?
  - How are the variables measured? Do they represent characteristics of what is measured?
  - Is additional data needed (e.g., larger sample, different or additional variables)?

- If data needs to be collected or gathered:
  - What variables need to be collected to address the issue or answer the question?
  - How can you measure these variables? Do they represent characteristics of what is measured?
  - How can data be collected? What are appropriate sources to gather data?
  - What methods can be used to collect or produce data (e.g., survey questions, measurements within an experiment)?
  - What sample design is appropriate?
  - How will data be accessed and stored?

**Consider issues with data.**

- Does the data come from a trustworthy source?
- What is your own personal connection to the data source?
- For what purpose, and in whose interest, was the data collected?
- Is there potential bias in the data collection or data source?
- Are there other ethical considerations about using a particular data source or collecting data needed to answer a question?

## Process the Data



**Consider strategies for processing and structuring data.**

- What strategies or techniques are most useful for obtaining or sourcing data?
- Where and how will data be stored and protected?
- Are there any issues with the ways data were entered? What will you do about possible erroneous/invalid data entries?
- What decisions will be made about missing data?
- What strategies or techniques will help process (e.g., clean messy data, organize, transform, etc.) and structure data in a consistent and usable format?
  - Which are the most efficient?
  - Which are the easiest to use?
  - Do you have the necessary skills and resources to carry these out?
- Should you merge multiple data tables or other structures?

**Consider processes that may help focus your investigation.**

- Do new cases need to be added to the data set?
- Is it helpful to sort, group or filter the data?
- Is it useful to create new variables based on the available variables?
- Do measurement units need to change?
- Is it useful to recode data values (e.g., change No/Yes to 0/1 to easily sum 1's)?

## Explore and Visualize Data

**Consider ways to visualize and summarize data.**

- How can data be modeled through visualizations (e.g., graphs, images, diagrams) to explore and reason about data in relation to the question(s) and context?
  - What visualizations draw attention to variability (spread) and other features of distributions (e.g., center, shape, outliers)?
  - What are affordances and drawbacks of different representations to make sense of data?
- How can data be modeled through statistical measures, and visualizations of these measures, to explore and reason about data in relation to the question(s) and context?
  - What statistical measures and visualizations of these measures will help to understand variability (spread) and other features of distributions (e.g., center, shape, outliers)?
  - What are affordances and drawbacks of different statistical measures to make sense of data?
- Are there any patterns or trends?
- Are there relationships amongst variables? If so, how can these relationships be described?

↕ **These two phases are interconnected** ↕

## Consider Models

**Investigate and select models.**

- What models are most useful in addressing the problem or questions (e.g., statistical measures, data visualizations, predictive models, distribution models)?
  - Which models summarize, describe, predict and/or explain variables and relationships between variables?
- What did you notice while exploring and visualizing the data (e.g., center, shapes, outliers, patterns, trends, relationships) that needs to be accounted for in selecting appropriate models to answer the problem?
- What aspects of the context and assumptions in data need to be considered when selecting models (e.g., Is mean or median more appropriate in the context? Do histograms or dot plots draw attention to key values in the context?)?
- How do the selected models account for and/or explain variability in the data?
- What are limitations of the models? How do those limitations need to be communicated and accounted for?

## Communicate and Propose Action

**Devise a strategy for communication.**

- What are the important issues within the problem context that stakeholders are interested in?
- Who is the audience? What information do they need to inform their decision-making?
- What are the best formats, media, and language for communicating findings and suggested actions?

**Develop and support your argument for claims and proposed actions.**

- How should you convey the problem, investigative question(s), methods, and analysis?
- Is it appropriate to discuss alternative approaches, models, or past results?
- What claims can be made from the data? What evidence is there to support these claims?
- What data visualizations could best support the claims? How are these visualizations interpreted? Do these visualizations need to be enhanced to be clearer to the audience?
- What statistical measures could best support the claims? How are these measures interpreted?
- Has uncertainty in findings been conveyed?
- What are the limitations, constraints, and potential biases of your data or analysis?
- What proposed actions within the context of the problem follow from the data investigation?
- What further data investigations should be recommended?
- Does the data story convey insights about the problem to your audience?

In today's world, most data investigations are supported by technological tools. The role of technology should be considered throughout each phase and many different technology tools may need to be utilized throughout an investigation, where each tool will likely have a different purpose. In Framing the Problem, internet-based tools (e.g., videos, images, reports, social media, maps, data dashboards) can help investigators understand, contextualize and situate a problem in a real-world phenomenon. At the Consider and Gather Data phase, one may need to consider issues related to what technology or other tools are available to help you collect data. For example, you may need to use survey and communication tools to design and invite participants to complete a survey, or perhaps use tools for data scraping, or web scraping, to import existing data from a website into a useful form (e.g., CSV file or spreadsheet file). You may use certain

tools to assist in processing the data (e.g., spreadsheets, Python or R), but other tools for exploring and visualizing the data and considering models (e.g., Tableau, CODAP, SAS, R, online applets). Finally, depending on your audience, you may use yet additional tools to communicate your findings and propose actions (e.g., video-makers, dashboard creators, word processing tools, presentation software). While teachers and students do not need to learn how to use all these tools before embarking on data investigations, there may be a need for learning just-in-time skills. In fact, teachers and students can learn how to use many tools through engaging in investigations.

For a broader description of the data investigation process, please see *The Data Investigation Process* resource. Available at: http://cdn.instepwithdata.org/DataInvestigationProcess.pdf

To cite this document:
Gemma F. Mojica, Hollylynne S. Lee, Emily Thrasher, Zachary Vaskalis, and Greg Ray. (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/ThinkingDataInvestigationProcess.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.

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.

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