Tools for reading and writing about data

2023 iSkills Workshop | Ciara Zogheib

Land Acknowledgement

We wish to acknowledge this land on which the University of Toronto operates. For thousands of years it has been the traditional land of the Huron-Wendat, the Seneca, and the Mississaugas of the Credit. Today, this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to have the opportunity to work on this land.

Hello!

- Ciara Zogheib (She/Her)
- PhD student at the iSchool
- Data scientist with the government

Stuff we need to do in the next 1.5 hours

Learn how to

- Better interpret data-centric research papers
- Examine the ways that data-focused methods and results are (and are not!) written about to better evaluate research
- Explore a selection of online resources and tools for communicating about data

WHO CARES

(Us!)

1 Process

What did we do with our data?

2 Words

What did we say about our data?

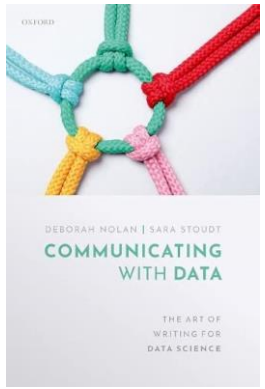
3 Pictures

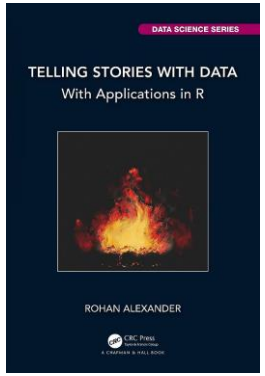
How do we display our data?

Resource

Nolan, D., & Stoudt, S. (2021). *Communicating with Data: The Art of Writing for Data Science.* Oxford University Press.

Alexander, R. (2023). *Telling stories with data: With applications in R (First edition).* CRC Press.





1

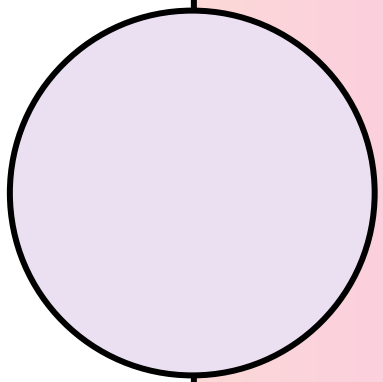
Process

What is ‘Data’?

- We can think of data as symbols representing observed properties of objects, events, or the world around us
- Data can be numbers, words, stories, colours, sounds - any systematically collected and organized observation
- Unstructured data = documents, images
- Structured data = data in tables, spreadsheets

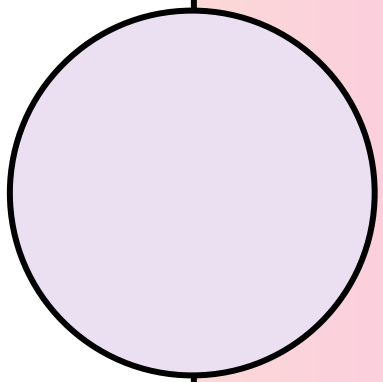
Data Collection

- **Setting**
  - Where were data collected?
  - When were data collected?
- **Sampling**
  - Probability sample?
  - Sample size
  - Recruitment (snowball, opportunistic, random)



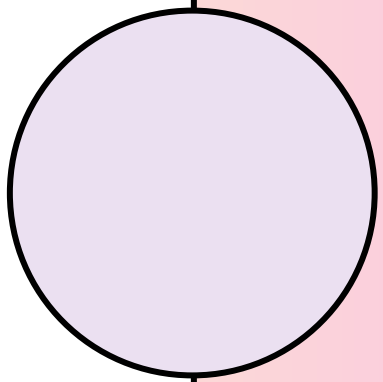
Data Collection

- **Inclusion/Exclusion Criteria**
  - What is and is not part of our dataset? How did we decide?
- **Missing Data**
  - Imputation
  - Complete case analysis



Data Collection

- **Variable Selection**
  - Are we measuring what we think we're measuring?
  - Justify proxies
- **Confounds**
  - Outside 'stuff' impacting our findings
  - Control or acknowledge as limitation



Data Analysis

Don’t panic when you see formal notation!

Resource

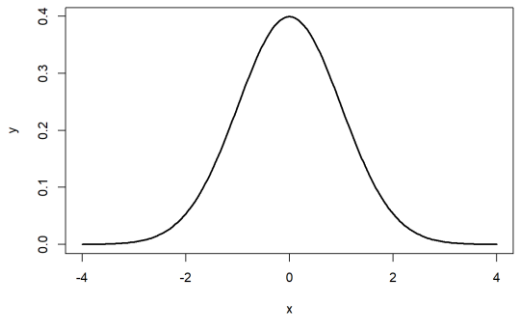
- Named equations?
- Look at methodological precedents (cited papers)
- https://www.statstutor.ac.uk/resources/uploaded/tutorsquickguidetostatistics.pdf
- https://menlocollege.libguides.com/OpenEducationResources/Statistics (open textbooks)
- https://libguides.uvic.ca/mathematics_statistics/oer (open textbooks)

Resource

- Named equations?
- Look at methodological precedents (cited papers)
- https://www.statstutor.ac.uk/resources/uploaded/tutorsquickguidetostatistics.pdf
- https://menlocollege.libguides.com/OpenEducationResources/Statistics (open textbooks)
- https://libguides.uvic.ca/mathematics_statistics/oer (open textbooks)

Data Analysis

- **What kind of stats?**
  - Descriptive (‘here’s what I have…’)
  - Inferential (‘here’s what this says about my phenomenon…’)
- **Assumptions**
  - Normality! Homoskedasticity!



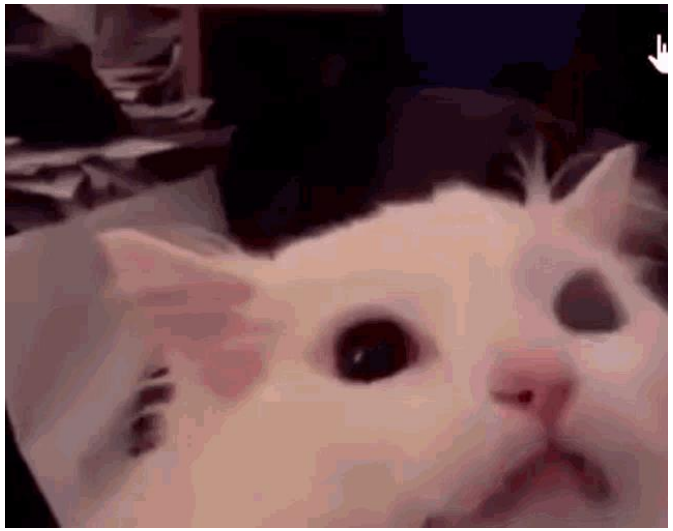
Data Analysis

- **Model Comparison**
  - How did they build this model?
  - How did they choose this model vs. others? (Accuracy? Fairness metrics? WAIC?)

https://seantrott.github.io/model_comparison/#High-level_goals

- **Reproducibility** 
  - Are data available?
  - Is code shared and commented?
  - Basically: could someone do it all again?

Data Analysis



The Dreaded p-Value

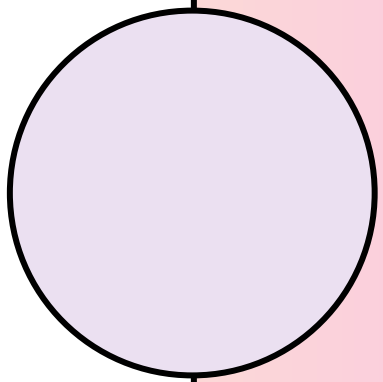
- p-values are useful for telling you a very specific thing
- No single value will give you the whole story

Ethics, Credit, Context

- **Who are the data about?**
  - Who did the analysis?
  - When were data collected?
- **Credit**
  - Who is doing the labour?
  - Who benefits? (Authorship, policy, funding, etc.)

Ethics, Credit, Context

- **Question Categories**
  - For our question or for convenience?
  - Who do we miss with binaries and groupings?
- **What are the data for?**
  - ‘The view from nowhere’
  - Own our intentions!



Resource

- Reflexivity and positionality are part of good data work!
- https://data-feminism.mitpress.mit.edu/ (Open-access textbook)
- https://blog.thegovlab.org/selected-readings-on-indigenous-data-sovereignty (Reading guide with links to other resources)

2

Words

Be Specific

- We can only make claims supported by our data and analyses
- Remember:

1. **Words mean things!** Colloquial vs. technical terms
2. **Define technical terms!**

Careful with Absolutes

- Avoid the trap of thinking that quant data = certainty in our results.
- Write honestly about limitations, conditions, or caveats to our processes.

Resource

- Academic Phrasebank
  https://www.phrasebank.manchester.ac.uk/
- Purdue Writing With Statistics
  https://owl.purdue.edu/owl/research_and_citation/using_research/writing_with_statistics/index.html
- ‘Making Data Talk’ Workbook
  https://www.cancer.gov/publications/health-communication/making-data-talk.pdf

But...

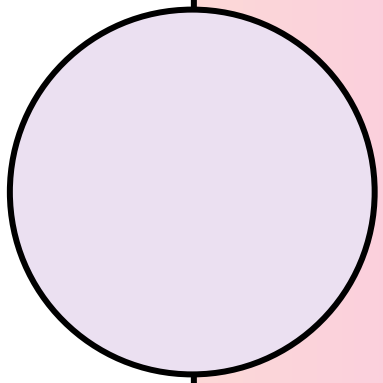
- **CONTEXT MATTERS!**
- Different disciplines and fields have different conventions for writing about data
  - Significant figures
  - Methodological Detail
  - Results Reporting
- Seek out examples!

3

Pictures

Picturing our Data

- **The basics**
  - Axes
  - Units and scales
  - Legends
  - Data labels
- **Chart type**
  - Intended function of our data
  - Structure of our data



Resource

- The Data Visualization Catalogue
  https://datavizcatalogue.com/

- Financial Times Visual Vocabulary
  https://community.powerbi.com/t5/Data-Stories-Gallery/FT-Visual-Vocabulary-Power-BI-Edition/td-p/584460

Accessibility

- Colour
  - Brightness and contrast (https://webaim.org/resources/contrastchecker/)
  - Palette (https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html)

Accessibility

- Font size
  - Context?
- Typeface
  - Serif? Sans-serif?
  - Variety?

Accessibility

- Alt text and description

Level	Semantic Content and Keywords
1	• Focus on chart elements such as chart type, title, axis ranges, labels, colours
2	• Descriptive statistics (eg. mean), outliers, max and min points, correlations, comparing points
3	• Complex trends, pattern synthesis, clusters, exceptions, common concepts • Perceptual interpretations of the data
4	• Context and domain insights, social and political context and explanations • Subjective interpretations that go beyond the data

Lundgard, A., & Satyanarayan, A. (2022). Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content.

1. “What is the dataset? **Who generated the dataset and why**?
2. **What is the process that underpins the dataset?** Given that process, **what is missing from the dataset or has been poorly measured?** Could other datasets have been generated, and if so, how different could they have been to the one that we have?
3. **What is the dataset trying to say, and how can we let it say this? What else could it say?** How do we decide between these?
4. **What are we hoping others will see from this dataset**, and how can we convince them of this? How much work must we do to convince them?
5. **Who is affected by the processes and outcomes, related to this dataset? To what extent are they represented in the dataset, and have they been involved in the analysis?**”

- Telling Stories with Data (Alexander, 2023)

Thank you!

More iSkills Workshops:
https://ischool.utoronto.ca/iskills/

Questions:
ciara.zogheib@mail.utoronto.ca