

Know Your Data / Know Your Tools

STAT 6106, Communicating Data and Statistics

Mitzi Morris

2023-09-11

Stan Development Team



Preamble: Bayesian Workflow (elements of)

Workflow is how you build a good model of the data.

- Exploratory data analysis / Data elicitation
- Model specification
- Model checking (simulated data)
 - Recover parameters, check prior predictive distribution, check calibration
- Model checking (simulated or observed data)
 - Check sample diagnostics, run posterior predictive checks
- Model comparison
 - Leave-one out cross-validation (LOO)

Preamble: Workflow in Context

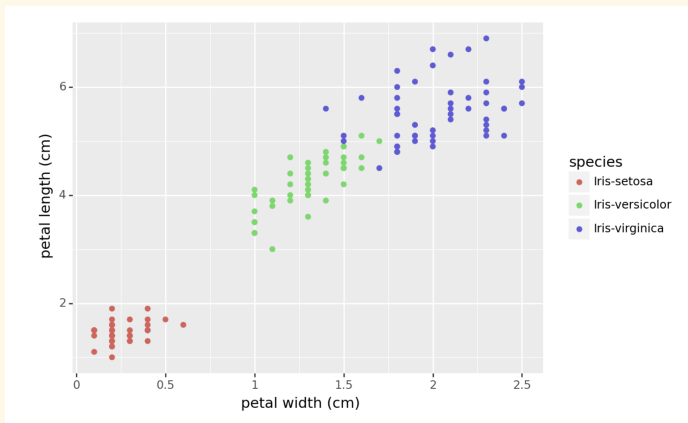
Your context determines what you communicate.

- Research priorities: innovation, pedagogy
 - Workflow is a blueprint for papers and lesson plans
- Applied concerns: accuracy, efficiency, generalizability
 - Workflow provides a principled choice between available models
- Challenge: building tools to automate workflow
 - Model choice
 - Model checking
 - ***Generating visualizations***

Challenges of Computing with Data

How much information is in your data?

- how many items? *for N items, only \sqrt{N} digits of accuracy*
- what is the resolution of the measurements? censored or truncated?



Challenges of Computing with Data

Floating-Point Standard: IEEE 754

- **Finite numbers** (s: sign; c: mantissa; q: exponent)

$$x = (-1)^s \times c \times 2^q$$

size	s, c bits	q bits	range	precision
32-bit	24	8	$\pm 3.4 \times 10^{38}$	7.2 digits
64-bit	53	11	$\pm 1.8 \times 10^{308}$	16 digits

- Quiet and signaling **not-a-number** (NaN)
- Positive and negative **infinity** ($+\infty, -\infty$)

Challenges of Computing with Data

Gaps Between Floating Point Numbers

- Smallest number greater than zero
 - single precision: 1.4×10^{-45}
 - double precision: 4.9×10^{-324}
- Largest number less than one
 - single precision: $1 - 10^{-7.2}$
 - double precision: $1 - 10^{-16}$
- Gap size *depends on scale*
 - Probabilities range from 0 - 1
 - The closer to zero, the more resolution

Challenges of Computing with Data

Catastrophic Cancellation

- Subtraction risks *catastrophic cancellation*
- Consider $0.99802 - 0.99801 = 0.00001$
 - input has five digits of precision
 - output has single digit of precision

E.g., problem for sample variance of sequence x

$$\text{var}(x) = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2$$

what happens to variance when all observations are close to the mean?

Challenges of Computing with Data

Takaway

- Understand computational pitfalls in order to detect/avoid them.
- Stan User's Guide
 - See *Part 2. Programming Techniques*
- Getting Started with Bayesian Statistics
 - See chapter *Making Stan Programs Go Faster*

Algebra for Data Science

Algebra

- sets, cartesian products
- graphs

Data




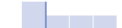


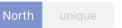
- Tabular data
- Structured data (XML, JSON, binary formats)

Processing Tabular Data

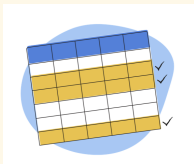
Background Reading: Data wrangling essentials

Data wrangling essentials: comparisons in JavaScript, Python, SQL, R, and Excel

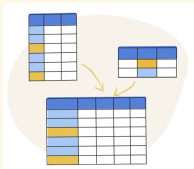
When adding JavaScript to your data work, it can be useful to see how it compares with other languages you've used before. Here, we show common data wrangling methods (like filtering, sorting, and adding columns) in JavaScript, Python, SQL, R, and Excel. All examples use the mock data below, stored as *salesData*:

▼	date date ▼ 	product string ▼ 	description string ▼ 	unitsSold integer ▼ 	unitPrice integer ▼ 	totalRevenue integer ▼ 	region 2 rows (40%) 
0	2022-01-01	Product A	Shirt - small	10	50	500	North
1	2022-01-02	Product B	Sweatshirt - large	15	25	375	South
2	2022-01-03	Product C	Jacket - small	8	75	600	West
3	2022-01-04	Product A	Shirt - medium	5	100	500	East
4	2022-01-05	Product B	Sweatshirt - large	20	30	600	North

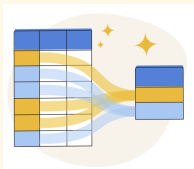
Processing Structured Data



Filtering rows in a table - same table structure, just fewer rows.



Joining two tables - many kinds of joins.

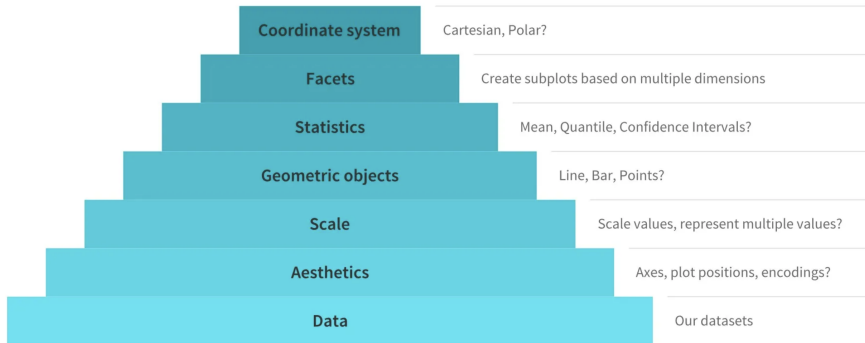


Grouping or aggregating information. Many ways to count.

Data Visualizations

Background Reading: Guide to the Grammar of Graphics

Major Components of the Grammar of Graphics



Grammar of Graphics: ggplot2 and plotnine

A grammar of graphics defines a plot in terms of:

- A dataset in the form of an R or pandas.DataFrame.
- A set of mappings from dataset variables to graph elements called “aesthetics”.
- A coordinate system, default Cartesian, x,y axes, where x is on the horizontal and y is on the vertical axis.
- A facet specification based on a categorical variable which results in per-category subplots, default None.
- One or more layers, each layer takes as arguments:
 - a dataset and aesthetic mapping - by default, the plot dataset and mappings are used
 - one geometric object (“geoms”)
 - one statistical transformation (“stats”), default “identity”
 - one position adjustment, default “identity”

Map-making

- If your data has significant geographic structure, show it!
- Both ggplot2 and plotnine can draw beautiful maps
- ***BUT!*** this requires getting the maps from a Geographic Information System(GIS)
- Let's do it! Jupyter notebook: Map-making with Plotnine

References

- Stan User's Guide
- Getting Started with Bayesian Statistics
- Data wrangling essentials
- Guide to the Grammar of Graphics
- Geographic Information System(GIS)
- Map-making with Plotnine

Questions?