

432 Class 01 Slides

thomaseLove.github.io/432

2022-01-11

Today's Agenda

- ➊ Mechanics of the course
- ➋ Why I write dates the way I do
- ➌ Data organization in spreadsheets
- ➍ Naming Things and Getting Organized
- ➎ Building a Table 1 (see Course Notes, Chapter 1)

There's a (pre-recorded) “live code” demo, too.

Welcome to 432.

Everything is at <https://thomaseLove.github.io/432/>

- Syllabus
- Calendar
 - with all deadlines, and links to class READMEs
- Course Notes
- Details on Assignments to come (+ next slide)
- R and Data
 - Updating / Installing R, RStudio, necessary R Packages
 - Review / Learn some R Basics (also see 431 web site)
- Sources
 - Books, Articles, YouTube series, etc.
- Links to Canvas, Piazza and Contact Us

Assignments

Every deliverable for the entire semester is listed in the Calendar, except for the Welcome to 432 Survey, which at least 30 of you've done, and if you haven't, visit <https://bit.ly/432-2022-welcome-survey>.

- Two projects
 - Project 1 (use publicly available data for linear & logistic models)
 - ① Proposal due 2022-01-31 (data selection, cleaning, exploration)
 - ② Final Materials due 2022-03-04 (analyses, discussion)
 - Project 2 (use almost any data you like and analyze it well)
- Two Quizzes (Quiz 1 due 2022-02-21, Quiz 2 due 2022-04-18)
 - Multiple choice and short answer, mostly, taken via a Google Form
- Six labs
 - Labs will be posted before our next class. Lab 01 is due Monday 2022-01-24 at 9 PM.
- Ten minute papers
 - First is due 2022-01-19. These actually take about 5 minutes each.

Syllabus and Instructions will provide more information on grading/feedback.

The Spring 2022 Teaching Assistants for 432 are:

- Stephanie Merlino Barr, PhD student in Clinical Translational Science
- Wyatt Bensken, PhD student in Epidemiology & Biostatistics
- Ali Elsharkawi, MS student in Clinical Research
- Shiyong Liu, PhD student in Epidemiology & Biostatistics
- Marie Michenkova, MS student in Biomedical Health Informatics
- Julia Yang Payne, PhD student in Clinical & Translational Science
- Monika Strah, MS student in Epidemiology & Biostatistics
- Yanning Wu, PhD student in Epidemiology & Biostatistics

All return from working with students in 431 this past Fall, and I couldn't be more grateful for their energy and effort. Learn more about the TAs in Section 6 of the Syllabus.

Getting Help

- Piazza is the location for discussion about the class. I follow it closely.
- We have 8 teaching assistants volunteering their time to help you.
- TAs will hold Office Hours beginning next Monday 2021-01-17 via Zoom, and the details will be available on Canvas (see Announcements) and our shared Google Drive.
- Dr. Love is available before and (especially) after class.
- Email Dr. Love directly only if you have a matter you need to discuss with him specifically. He's at Thomas dot Love at case dot edu.

We WELCOME your questions/comments/corrections/thoughts!

Tools You Will Definitely Use in this Class

- **Course Website** (see the bottom of this slide) especially the Calendar
 - Each class has a README (announcements, reminders, etc.) plus slides
- **R, RStudio and R Markdown** for, well, everything
- **Canvas** for access to Zoom meetings *and 432 recordings*, submission of most assignments
- **Google Drive via CWRU** for *recordings from 431*, forms (Minute Papers/Surveys/Quizzes) and receiving feedback on labs, projects, and Minute Papers
- **Piazza** is our discussion board. It's a moderated place to ask questions, answer questions of your colleagues, and get help fast. You don't have to pay to use it.
- **Zoom** for class sessions and for TA office hours

Some source materials are **password-protected**. What is the password?



An approximate answer to the right
problem is worth a good deal more
than an exact answer to an
approximate problem.

— *John Tukey* —

AZ QUOTES

How To Write Dates (https://xkcd.com/1179/)


PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013.II.27. $27\frac{1}{2}$ -13 2013.158904109
MMXIII-II-XXVII MMXIII $\frac{LVII}{CCCLXV}$ 1330300800
 $((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$ 2013
10/11011/1101 02/27/20/13 $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ & 5 & 6 & 7 & 8 \end{matrix}$ 

Tidy Data (Wickham)

“A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. . . .

Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table.

This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.”

<https://www.jstatsoft.org/article/view/v059i10>

“Data Tidying” presentation in *R for Data Science*

- Defines tidy data
- Demonstrates methods for tidying messy data in R

Read Sections

- 5 (Data transformation),
- 10 (Tibbles), 11 (Data import) and, especially, 12 (Tidy data)

<https://r4ds.had.co.nz/>

Data Organization in Spreadsheets (Broman & Woo)

- Create a data dictionary.
 - Jeff Leek has good thoughts on this in “How to Share Data with a Statistician” at <https://github.com/jtleek/datasharing>
 - Shannon Ellis and Jeff Leek’s preprint “How to Share data for Collaboration” touches on many of the same points at <https://peerj.com/preprints/3139v5.pdf>

We want:

- 1 The raw data.
- 2 A tidy data set.
- 3 A codebook describing each variable and its values in the tidy data set.
- 4 An explicit and exact recipe describing how you went from 1 to 2 and 3.

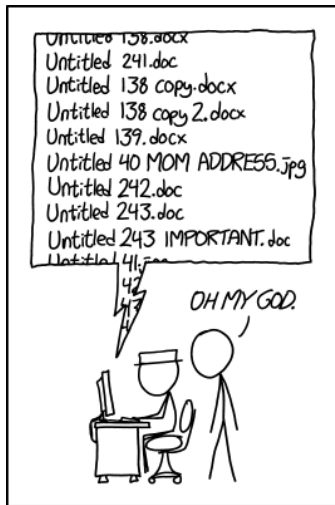
Data Organization in Spreadsheets: Be Consistent

- Consistent codes for categorical variables.
 - Either “M” or “Male” but not both at the same time.
 - Make it clear enough to reduce dependence on a codebook.
 - No spaces or special characters other than `_` in category names.
- Consistent fixed codes for missing values.
 - NA is the most convenient R choice.
- Consistent variable names
 - In R, I'll use `clean_names` from the `janitor` package to turn everything into `snake_case`.
 - In R, start your variable names with letters. No spaces, no special characters other than `_`.
- Consistent subject / record identifiers
 - And if you're building a `.csv` in Excel, don't use ID as the name of that identifier.
- Consistent data layouts across multiple files.

What Goes in a Cell?

- Make your data a rectangle.
 - Each row represents a record (sometimes a subject).
 - Each column represents a variable.
 - First column is a unique identifier for each record.
- No empty cells.
- One Thing in each cell.
- No calculations in the raw data
- No font colors
- No highlighting

Naming Files is Hard (<https://xkcd.com/1459/>)



PROTIP: NEVER LOOK IN SOMEONE
ELSE'S DOCUMENTS FOLDER.

NO

myabstract.docx

Joe's Filenames Use Spaces and Punctuation.xlsx

figure 1.png

fig 2.png

JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt

YES

2014-06-08_abstract-for-sla.docx

joes-filenames-are-getting-better.xlsx

fig01_scatterplot-talk-length-vs-interest.png

fig02_histogram-talk-attendance.png

1986-01-28_raw-data-from-challenger-o-rings.txt

Data Organization in Spreadsheets: Use consistent, strong file names.

Jenny Bryan's advice on "Naming Things" hold up well. There's a full presentation at SpeakerDeck.

Good file names:

- are machine readable (easy to search, easy to extract info from names)
- are human readable (name contains content information, so it's easy to figure out what something is based on its name)
- play well with default ordering (something numeric first, left padded with zeros as needed, use ISO 8601 standard for dates)

Avoid: spaces, punctuation, accented characters, case sensitivity

left pad other numbers with zeros

```
01_marshall-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

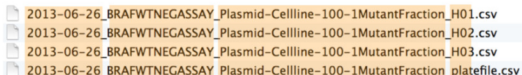
if you don't left pad, you get this:

```
10_final-figs-for-publication.R
1_data-cleaning.R
2_fit-model.R
```

which is just sad

Jenny Bryan: Deliberate Use of Delimiters

Deliberately use delimiters to make things easy to compute on and make it easy to recover meta-data from the filenames.



2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv

```
> flist <- list.files(pattern = "Plasmid") %>% head

> stringr::str_split_fixed(flist, "[_\\.]", 5)
      [,1]      [,2]      [,3]      [,4] [,5]
[1,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A01" "csv"
[2,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A02" "csv"
[3,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A03" "csv"
[4,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B01" "csv"
[5,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B02" "csv"
[6,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B03" "csv"
```

“_” underscore used to delimit units of meta-data I want later

“-” hyphen used to delimit words so my eyes don’t bleed

Don't get too cute.



Jenny Bryan

@JennyBryan

Following



The Golden Rule of Naming Files and Other Things:

Thou shalt get only as creative with names as thy own skill with regular expressions.

11:31 PM - 10 Dec 2016

Goal: Avoid this...

SIMPLY EXPLAINED



Idea from Jen Simmons and John Albin Wilkins during episode #40 of "Web Ahead" about Git:
<http://5by5tv/webahead/40>

Be organized

do this as you go, not "tomorrow"

but also don't fret over past mistakes
raise the bar for *new* work

Don't spend a lot of time bemoaning or cleaning up past ills. Strive to improve this sort of thing going forward.

“Good Enough Practices in Scientific Computing”

- 1 Save the raw data.
- 2 Ensure that raw data is backed up in more than one location.
- 3 Create the data you wish to see in the world (the data you wish you had received.)
- 4 Create analysis-friendly, tidy data.
- 5 Record all of the steps used to process data.
- 6 Anticipate the need for multiple tables, and use a unique identifier for every record.

<http://bit.ly/good-enuff>

Lots of great advice here on software, collaboration and project organization.

Something Practical: Building Table 1



January 18, 2019

Incidence, Risk Factors, and Outcomes Associated With In-Hospital Acute Myocardial Infarction

Steven M. Bradley, MD, MPH^{1,2}; Joleen A. Borgerding, MS³; G. Blake Wood, MS³; [et al](#)

» [Author Affiliations](#) | [Article Information](#)

JAMA Netw Open. 2019;2(1):e187348. doi:10.1001/jamanetworkopen.2018.7348

Key Points

Question What are the incidence, risk factors, and outcomes associated with in-hospital acute myocardial infarction (AMI)?

Findings This cohort study of 1.3 million patients hospitalized in US Veterans Health Administration facilities found an incidence of in-hospital AMI of 4.27 per 1000 admissions, and risk factors associated with in-hospital AMI included history of coronary artery disease, elevated heart rate, low hemoglobin level, and elevated white blood cell count. Compared with a matched control group, mortality was significantly higher for in-hospital AMI.

Meaning In-hospital AMI is common and is associated with prior cardiovascular disease, physiological disturbances, and poor survival.

Part of Bradley et al.'s Table 1

Table 1. Patient Characteristics on Admission and In-Hospital Variables Prior to Event for Matched In-Hospital Acute Myocardial Infarction Cases and Controls

Characteristic	No. (%)			P Value
	Total (N = 1374)	Cases (n = 687)	Controls (n = 687)	
Age, mean (SD), y	73.3 (10.2)	73.3 (10.1)	73.4 (10.3)	.80
Male	1343 (97.7)	677 (98.5)	666 (96.9)	.05
White race/ethnicity	1073 (78.1)	546 (79.5)	527 (76.7)	.22
Married	666 (48.5)	356 (51.8)	310 (45.1)	.01
Location				
Intensive care unit	251 (18.3)	186 (27.1)	65 (9.5)	<.001
Medical bed	1026 (74.7)	446 (64.9)	580 (84.4)	
Other	97 (7.1)	55 (8.0)	42 (6.1)	

Table Creation Instructions, JAMA: linked here

Creating a Table

Use the table editor of the word processing software to build a table. Do not embed tables as images in the manuscript file or upload tables in image formats. Regardless of which program is used, each piece of data needs to be contained in its own cell in the table. Tables should be single-spaced.

Avoid creating tables using spaces or tabs. For accepted manuscripts, tables created with spaces, tabs, and/or hard returns must be retyped during the editing process, creating delays and opportunities for error. Do not try to align cells with hard returns or extra spaces. Similarly, no cell should contain a hard return or tab. Although individual empty cells are acceptable in a table, be sure there are no empty columns.

Place each row of data in a separate row of cells:

Table 1. Title

Treatment	Group A	Group B
Medical	500	510
Surgical	500	490

Note that numbers and percentages are presented in the same cell, and measures of variability are in the same cell as their corresponding statistic:

Table 2. Title

Characteristics	Group A (n = 50)	Group B (n = 50)	Relative Risk (95% CI)
Women, No. (%)	25 (50)	20 (40)	1.25 (1.11-1.57)
Age, mean (SD), y	35 (8)	37 (7)	0.98 (0.92-1.05)

the rows. In Table 3, the final column lists the *P* value for the overall age comparison:

Table 3. Title

Age, y	Blood Pressure, mm Hg	<i>P</i> Value
18-34	120/75	
35-50	110/80	.08
51-80	125/82	

The table should be constructed such that comparisons between groups read horizontally (see Tables 1 and 2).

Do not draw lines or rules—the table grid feature will display the outlines of each cell.

Data Presentation

When presenting percentages, include numbers (numerator, and denominator if necessary). Include variability where applicable (eg, mean [SD] or median [interquartile range]).

All *P* values should be reported as exact numbers to 2 digits past the decimal point, regardless of significance, unless they are lower than .01, in which case they should be presented to 3 digits. Express any *P* values lower than .001 as $P < .001$. *P* values can never equal 0 or 1.

Footnotes

Be sure to explain empty cells. Also, if necessary add a footnote to explain why numbers may not sum to group totals or percentages do not total 100. List abbreviations for the table in a footnote and use superscript letters to mark each footnote (a,b,c, etc).

Questions

For questions on table construction or formatting, contact Stacy Christian, Director of Formatting and Design, at stacy.christian@jamanetwork.com.

A Data Set

The `bradley.csv` data set on our web site is simulated, but consists of 1,374 observations (687 Cases and 687 Controls) containing:

- a subject identification code, in `subject`
- `status` (case or control)
- `age` (in years)
- `sex` (Male or Female)
- `race/ethnicity` (white or non-white)
- `married` (1 = yes or 0 = no)
- `location` (ICU, bed, other)

The `bradley.csv` data closely match the summary statistics provided in Table 1 of the Bradley et al. article. Our job is to recreate that part of Table 1, as best as we can.

The bradley.csv data (first 5 rows)

- The bradley_sim.md file on our web site shows you how I simulated the data.

	A	B	C	D	E	F	G
1	subject	status	age	sex	race_eth	married	location
2	1	Control	64	Male	white	1	Bed
3	2	Case	70	Male	white	1	ICU
4	3	Control	68	Male	white	0	Bed
5	4	Control	76	Male	white	1	Bed
6	5	Control	70	Male	white	1	Bed

To “Live” Coding

On our web site (Data and Code + Class 01 materials)

- In the data folder:
 - `bradley.csv` data file
- `bradley_table1.Rmd` R Markdown script
- `bradley_table1.md` Results of running R Markdown
- `bradley_table1_result.csv` is the table generated by that R Markdown script

To The “Live Code”

Opening bradley_table1_result.csv in Excel

	A	B	C	D	E	
1		Case	Control	p	test	
2	n	687	687			
3	age (mean (SD))	73.78 (10.24)	72.60 (10.50)	0.035		
4	sex = Male (%)	677 (98.5)	666 (96.9)	0.069		
5	race_eth = white (%)	546 (79.5)	527 (76.7)	0.24		
6	marital = yes (%)	356 (51.8)	310 (45.1)	0.015		
7	loc (%)			<0.001		
8	ICU	186 (27.1)	65 (9.5)			
9	Bed	446 (64.9)	580 (84.4)			
10	Other	55 (8.0)	42 (6.1)			
11						

Learning More About Table 1

Chapter 1 of the Course Notes covers two larger examples, and more details, like. . .

- specifying factors, and re-ordering them when necessary
- using non-normal summaries or exact categorical tests
- dealing with warning messages and with missing data
- producing Table 1 in R so you can cut and paste it into Excel or Word

FYI: Lab 01 requires you to build a Table 1 from data.

Wrapping Up

Today we discussed

- 1 Why I write dates the way I do
- 2 Mechanics of the course
- 3 Data organization in spreadsheets
- 4 Naming Things and Getting Organized
- 5 Building a Table 1 (review Course Notes, Chapter 1)

Next Steps?

- Complete the Welcome to 432 survey by tomorrow noon.
- Look over the syllabus and the website, get R up and running.
- Look at the suggestions in the Class 01 README.
- Get started reading Leek's short book.
- You **can** do this.