

432 Class 1 Slides

github.com/THOMASELOVE/2019-432

2019-01-22

Welcome to 432.

Everything is at <https://github.com/THOMASELOVE/2019-432>

- Syllabus
- Calendar
- Slides
- Data and Code
- List of R Packages
- References
- Deliverables
 - 6 Homework assignments, the first is due 2019-02-01
 - 2 Projects, the first proposal is due 2019-02-15
 - 2 Quizzes, the first of which you'll get on 2019-03-01.
 - Minute Papers, the first of which you'll get today.
 - Class Participation

Getting Help

- Email 431-help at case dot edu 24/7 through May 7.
- Dr. Love is available before and (especially) after class.
- Course Calendar and Syllabus tell you more about TAs and deadlines. Most assignments (homework and project proposals) are due on Fridays at 2 PM.
- Most things are submitted through Canvas or through Google Forms

TAs this semester are Bob Winkelman, Satyakam Mishra, Maher Kazimi, Zuxi (Terry) Cui and Xueyi (Julia) Zhang. Learn more about them in Section 3 of the Syllabus.

TA Office Hours are held in WG-56 and WG-67

- Monday - Friday 11:30 AM to 12:45 PM
- Tuesday and Thursday 2:30 PM to 3:45 PM, as well.

Broman and Woo, “Data Organization in Spreadsheets”

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.

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 ^{LVII}_{CCCLXV} 1330300800

$((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$ ~~2013~~ Miss 

10/11011/1101 02/27/20/13 $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ & 5 & 6 & 7 & 8 \end{matrix}$

Data Organization in Spreadsheets

- 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.

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

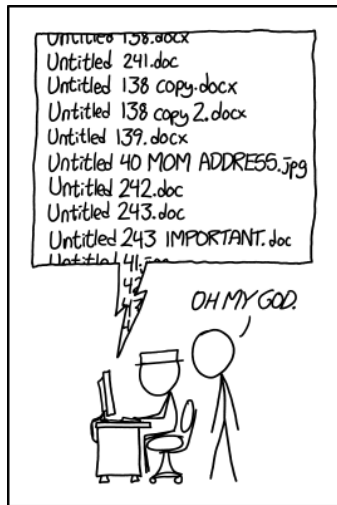
- 5 (Data transformation),
- 10 (Tibbles), 11 (Data import) and, especially, 12 (Tidy data)

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

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/>)



PRO TIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

How To Name Files

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

from Jenny Bryan's "Naming Things" slides...

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
```

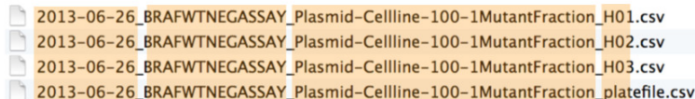
if you don't left pad, you get this:

```
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



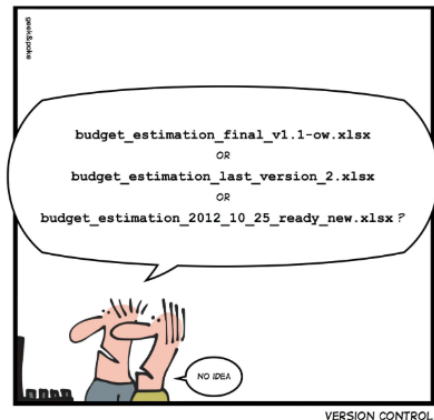
The Golden Rule of Naming Files and Other Things:

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”

Good enough practices in scientific computing

Wilson, Bryan, Cranston, Kitzes, Nederbragt, Teal

<https://doi.org/10.1371/journal.pcbi.1005510>

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



From “Good Enough”

- 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

A New Original Investigation

Original Investigation | Cardiology



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.

Link to Source

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)

To present data that span more than 1 row, do not merge the cells vertically. Instead, put the data in a cell near the middle of

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 Christiansen, director of manuscript editing, at stacy.christiansen@jama-archives.org.

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	Case	106	Male	non-white	1	ICU
3	2	Case	71	Female	white	1	Other
4	3	Control	67	Male	white	1	Bed
5	4	Case	68	Male	white	0	Bed
6	5	Control	57	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))	73.20 (9.96)	73.16 (9.75)	0.941	
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)		

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

Homework 1 also requires you to build a Table 1 from data.

Wrapping Up

Today we discussed

- 1 Data organization in spreadsheets
- 2 Building a Table 1 (review Course Notes, Chapter 1)

Thursday, we'll discuss:

- 1 Frank Harrell's Directory of Statistical Terms
- 2 BRFSS/SMART example (from Course Notes, Chapter 2)
- 3 Minute Paper Results

Minute Paper due tomorrow (2019-01-23) at 2 PM

Please visit the link provided on the Class 01 README to find the Google Form where you'll answer a few questions.