

432 Class 01

<https://thomaseLove.github.io/432-2024/>

2024-01-16

Getting To These Slides

Our web site: <https://thomaseLove.github.io/432-2024/>

- Note that this link is posted to the bottom of every slide.

Visit the Calendar at the top of the page, which will take you to the Class 01 README page.

- These Slides for Class 01 are linked at the Class 01 README.
 - We'll look at the **HTML slides** during class.
 - We also provide the Quarto code I used to build the slides.

Today's Agenda

- ① Mechanics of the course
- ② Why I write dates the way I do
- ③ Data organization in spreadsheets
- ④ Naming Things and Getting Organized
- ⑤ Switching from R Markdown to Quarto
- ⑥ Building and Validating small models for Penguin Bill Length

Section 1

Course Mechanics

Welcome to 432.

Just about everything is linked at <https://thomaselove.github.io/432-2024>

- Calendar
 - final word on all deadlines, and links to each class and TA office hours.
- Syllabus (can download as PDF)
- Course Notes HTML and PDF

Also linked on our website

- Software
 - Updating / Installing R and RStudio, necessary R Packages
- Get Data (Code, Quarto templates) at our 432-data page
- Assignments (Labs, Projects, Quizzes - see next few slides)
- Sources (books, articles, videos, etc.)
- Key Links (Canvas, Campuswire, Shared Google Drive, Minute Papers)
- Contact Us (Campuswire + TA office hours + My email)

Assignments

Every deliverable is listed in the Calendar.

- Welcome to 432 Survey at <https://bit.ly/432-2024-welcome-survey> due tomorrow (Wednesday 2024-01-17) at Noon.
 - Be sure you see the course in Campuswire and on Canvas. Thanks.

Assignments include two projects, eight labs, ten minute papers and two quizzes.

Two Projects

Project A (publicly available data: linear & logistic models)

- 1 Plan: February 12 (data selection, cleaning, exploration)
- 2 Final Portfolio & (recorded) Presentation due March 18

Project B (use almost any data and build specific models)

- 1 Proposal Form April 10
- 2 Presentation (in-person or Zoom) in late April / early May
- 3 Portfolio (prepared using Quarto) due May 7

Eight Labs

Eight labs, meant to be (generally) shorter than 431 Labs

- ❶ Lab 1 is due Tuesday 2024-01-23 at Noon.
- ❷ Lab 2 is due Tuesday 2024-01-30 at Noon.
- Instructions are available now for all 8 labs. First 7 due Tuesdays at Noon.

Lab 8 can be done at any time, and involves building (or augmenting) a website for yourself.

Two Quizzes, Ten Minute Papers

- Quiz 1 in late February, Quiz 2 in late April
 - Receive Quiz on Thursday at 5 PM, due Tuesday at Noon.
 - Mostly multiple choice or short answer, via Google Form.
- Minute Papers (due most Wednesdays at Noon.)
 - First is after Class 3 due 2024-01-24.
 - About 5 minutes each, also done via Google Form

Syllabus, Lab Instructions provide feedback details.

Getting Help

- Campuswire is the location for discussion about the class.
- 7 teaching assistants volunteering their time to help you.
- TAs hold Zoom Office Hours starting Friday 2024-01-19.
- Dr. Love is also available after every class to chat.
- Email Dr. Love if you have a matter you need to discuss with him, at **Thomas dot Love at case dot edu.**

We WELCOME questions/comments/corrections/thoughts!

Spring 2024 432 Teaching Assistants

- Ali Elsharkawi, MS student in Clinical Research
- Chenyu Liu, PhD student in Epidemiology & Biostatistics
- Alex Olejko, PhD candidate in Psychological Sciences
- Lindsay Petrenchik, PGY1 Pharmacy Resident at UH Samaritan Medical Center
- Miza Salim Hammoud, MS in Clinical Research, postdoc at Cleveland Clinic
- Monika Strah, PhD student in Clinical Translational Science
- Ria Tilve, MD student and MPH graduate

Spring 2024 432 Teaching Assistants

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 the Syllabus.

TA Zoom Office Hours begin this Friday 2024-01-19. Details coming soon to Canvas, our website, and our Shared Google Drive.

Tools You Will Use in this Class

- **Course Website** (bottom of every slide) especially the Calendar
 - Each class has a README plus slides
- **R, RStudio and Quarto** for, well, everything
- **Canvas** for access to Zoom meetings and 432 recordings, submission of Labs and Project assignments
- **Google Drive via CWRU** for forms (Minute Papers/Surveys/Quizzes) and for feedback on assignments

Tools You Will Use in this Class

- **Campuswire** is our discussion board. It's a moderated place to ask questions, answer questions of your colleagues, and get help fast. Open 24/7.
- **Zoom** for class recordings and 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

Section 2

Why I Write Dates The Way I Do

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~~  Missss

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

Section 3

Data Organization in Spreadsheets

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.

Tidy Data (continued)

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*, 2e

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

Read Sections 3 (Data transformation) and 5 (Data tidying)

<https://r4ds.hadley.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>

Sharing Data with a Statistician

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.

Data Organization in Spreadsheets: **Be Consistent**

- 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 `_`.

Data Organization in Spreadsheets: **Be Consistent**

- 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.

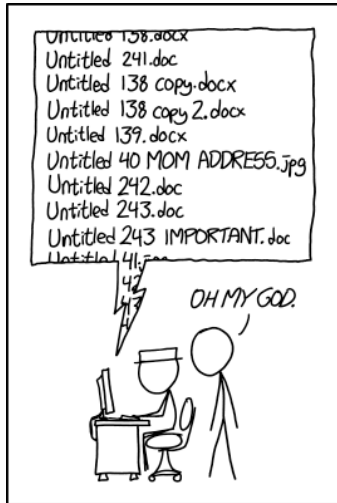
What Goes in a Cell?

- 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 and no highlighting

Section 4

Naming Things and Getting Organized

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)

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

Good file names:

- 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...

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
```

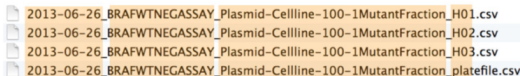
if you don't left pad, you get this:

```
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



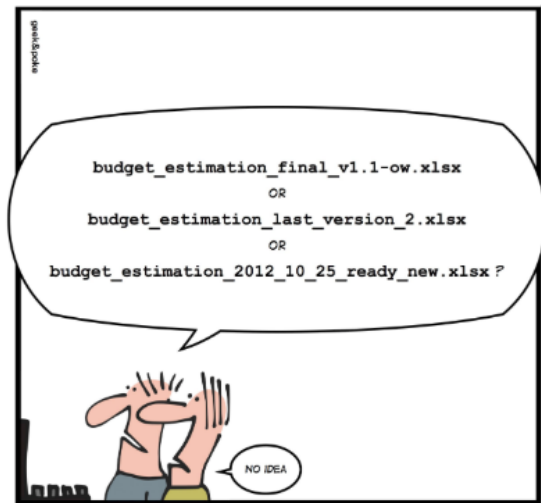
The Golden Rule of Naming Files and Other Things:

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



VERSION CONTROL

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”

- 1 Save the raw data.
- 2 Ensure that raw data is backed up more than once.
- 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.

Section 5

Switching from R Markdown to Quarto

R Markdown to Quarto

<https://quarto.org/> is the main website for Quarto.

If you can write an R Markdown file, it will also work in Quarto, by switching the extension from `.rmd` to `.qmd`.

- We provide a Quarto template for Lab 1 (due Tuesday at Noon) which should ease your transition a bit.
- Read Chapter 30 (Quarto) in R for Data Science, 2e
- Lots of other suggestions in the Class 01 README and our Sources page.

All material for this course is written using Quarto.

Section 6

Building and Validating Linear Prediction Models

R Setup

```
knitr::opts_chunk$set(comment = NA)

library(broom); library(glue); library(gt)
library(janitor); library(knitr); library(mosaic)
library(patchwork); library(rsample)
library(palmerpenguins); library(tidyverse)

theme_set(theme_bw())
```

Data Load

```
our_tibble <- penguins |>
  select(species, sex, bill_length_mm) |>
  drop_na()

our_tibble |> summary()
```

species	sex	bill_length_mm
Adelie :146	female:165	Min. :32.10
Chinstrap: 68	male :168	1st Qu.:39.50
Gentoo :119		Median :44.50
		Mean :43.99
		3rd Qu.:48.60
		Max. :59.60

Partition our `tibble` into training/test samples

We will place 60% of the penguins in our training sample, and require that similar fractions of each species occur in our training and testing samples. We use functions from the **rsample** package here.

```
set.seed(20240117)
our_split <- initial_split(our_tibble, prop = 0.6,
                           strata = species)
our_train <- training(our_split)
our_test  <- testing(our_split)
```

We could use `slice_sample()` as in the Course Notes if we didn't stratify by species.

Result of our partitioning

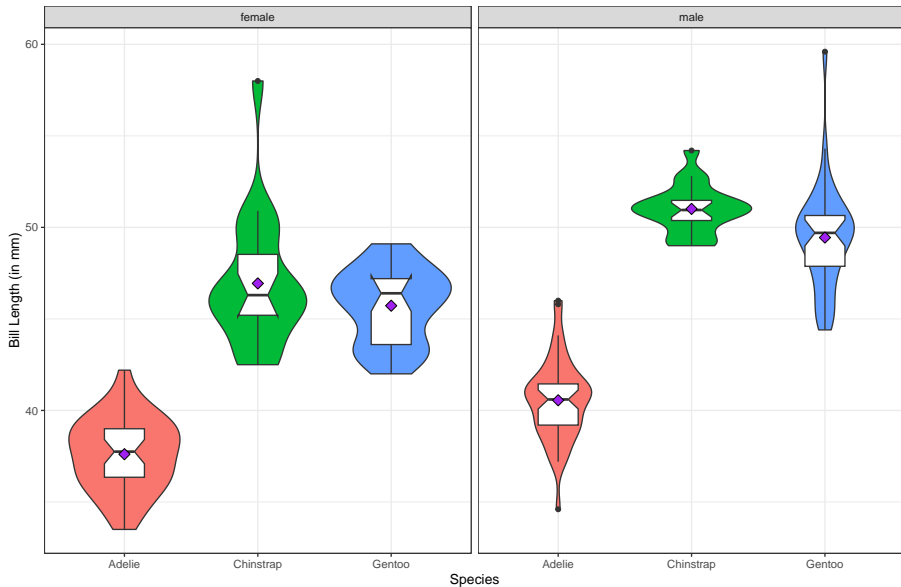
```
our_train |> tabyl(species) |> adorn_totals() |>  
  adorn_pct_formatting()
```

species	n	percent
Adelie	87	43.9%
Chinstrap	40	20.2%
Gentoo	71	35.9%
Total	198	100.0%

```
our_test |> tabyl(species) |> adorn_totals() |>  
  adorn_pct_formatting()
```

species	n	percent
Adelie	59	43.7%
Chinstrap	28	20.7%
Gentoo	48	35.6%
Total	135	100.0%

Bill Length, by Species, faceted by Sex
198 of the Palmer Penguins



Code for previous slide

```
ggplot(data = our_train,
       aes(x = species, y = bill_length_mm)) +
  geom_violin(aes(fill = species)) +
  geom_boxplot(width = 0.3, notch = TRUE) +
  stat_summary(fill = "purple", fun = "mean",
              geom = "point",
              shape = 23, size = 3) +
  facet_wrap(~ sex) +
  guides(fill = "none") +
  labs(title = "Bill Length, by Species, faceted by Sex",
       subtitle =
         glue(nrow(our_train), " of the Palmer Penguins"),
       x = "Species", y = "Bill Length (in mm)")
```


Model m1

```
m1 <- lm(bill_length_mm ~ species + sex, data = our_train)

anova(m1)
```

Analysis of Variance Table

Response: bill_length_mm

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
species	2	3962.6	1981.32	353.74	< 2.2e-16 ***
sex	1	587.3	587.31	104.86	< 2.2e-16 ***
Residuals	194	1086.6	5.60		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model 1 coefficients

```
tidy(m1, conf.int = TRUE, conf.level = 0.90) |>  
  select(term, estimate, conf.low, conf.high) |>  
  kable(digits = 1)
```

term	estimate	conf.low	conf.high
(Intercept)	37.3	36.8	37.9
speciesChinstrap	9.9	9.2	10.7
speciesGentoo	8.5	7.9	9.2
sexmale	3.5	2.9	4.0

Model m2

```
m2 <- lm(bill_length_mm ~ species, data = our_train)

## anova(m2) yields p-value < 2.2e-16 (not shown here)

tidy(m2, conf.int = TRUE, conf.level = 0.90) |>
  select(term, estimate, conf.low, conf.high) |>
  kable(digits = 1)
```

term	estimate	conf.low	conf.high
(Intercept)	39.2	38.7	39.7
speciesChinstrap	9.8	8.8	10.7
speciesGentoo	8.5	7.7	9.3

In-Sample Comparison

```
bind_rows(glance(m1), glance(m2)) |>
  mutate(model = c("m1 (species & sex)",
                    "m2 (species only)")) |>
  select(model, r2 = r.squared, adjr2 = adj.r.squared,
         AIC, BIC, sigma, nobs) |>
  kable(digits = c(0, 3, 3, 1, 1, 2, 0))
```

model	r2	adjr2	AIC	BIC	sigma	nobs
m1 (species & sex)	0.807	0.804	909.0	925.4	2.37	198
m2 (species only)	0.703	0.700	992.6	1005.7	2.93	198

Which model has better in-sample performance?

Assessing Performance in Test Sample

```
m1_aug <- augment(m1, newdata = our_test)

m1_res <- m1_aug |>
  summarize(val_R_sq = cor(bill_length_mm, .fitted)^2,
            MAPE = mean(abs(.resid)),
            RMSPE = sqrt(mean(.resid^2)),
            max_Error = max(abs(.resid)))

m2_aug <- augment(m2, newdata = our_test)

m2_res <- m2_aug |>
  summarize(val_R_sq = cor(bill_length_mm, .fitted)^2,
            MAPE = mean(abs(.resid)),
            RMSPE = sqrt(mean(.resid^2)),
            max_Error = max(abs(.resid)))
```

Test Sample Performance

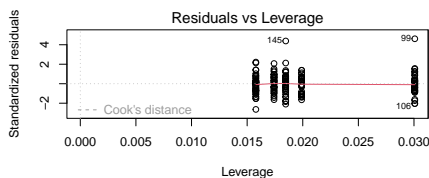
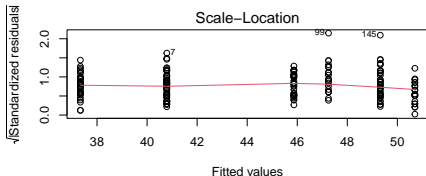
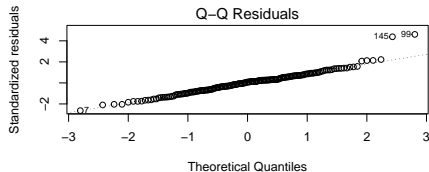
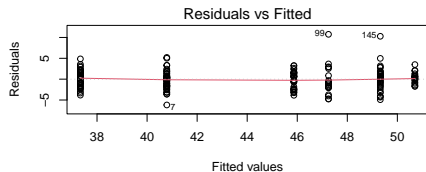
```
bind_rows(m1_res, m2_res) |>
  mutate(model = c("m1 (species & sex)",
                    "m2 (species only)")) |>
  relocate(model) |>
  kable(digits = c(0, 3, 2, 2, 1))
```

model	val_R_sq	MAPE	RMSPE	max_Error
m1 (species & sex)	0.841	1.75	2.29	6.6
m2 (species only)	0.718	2.54	3.06	8.2

Which model predicts better in the test sample?

Residual Plots for m1 (training)

```
par(mfrow = c(2,2)); plot(m1); par(mfrow = c(1,1))
```



What we did in this example...

- ① R packages, usual commands, ingest the data.
- ② Look at what we have and ensure it makes sense. (DTDP)
- ③ Partition the data into a training sample and a test sample.
- ④ Run a two-way ANOVA model (called `m1`) in the training sample; evaluate the quality of fit.
- ⑤ Run a one-way ANOVA model (called `m2`) in the training sample; evaluate the quality of fit.

What we did in this example...

- ⑥ Use augment to predict from each model into the test sample; summarize and compare predictive quality.
- ⑦ Choose between the models and evaluate assumptions for our choice.

For Next Time...

- ❶ If you're not registered with SIS, do so, for PQHS/CRSP/MPHP 432.
- ❷ Review the website and Calendar, and skim the Syllabus and Course Notes.
- ❸ Welcome to 432 Survey at <https://bit.ly/432-2024-welcome-survey> by noon Wednesday 2024-01-17.

For Next Time...

- ④ Accept the invitation to join the Campuswire Discussion Forum for 432.
- ⑤ Buy Jeff Leek's How to be a Modern Scientist and read it by the end of January.
- ⑥ Get started installing or updating the software you need for the course.
- ⑦ Get started on Lab 1, due Tuesday 2024-01-23 at Noon.