Agenda - DSC 482 / MTH 445

Agenda - DSC 482 - Applied Math for Data Science, Fall 2022

README

- DSC 482 focuses on statistics, probability and distributions
- We use R as a language, GNU Emacs as editor, ESS as statistics extension, and Org-mode for literate programming
- For details on objectives, audience, grading, schedule, check the syllabus or the FAQ on GitHub.
- This file contains a (dynamically updated) agenda for each session as well as some content.
- You can also look at the agenda on GitHub.

FAQ - Frequently Asked Questions

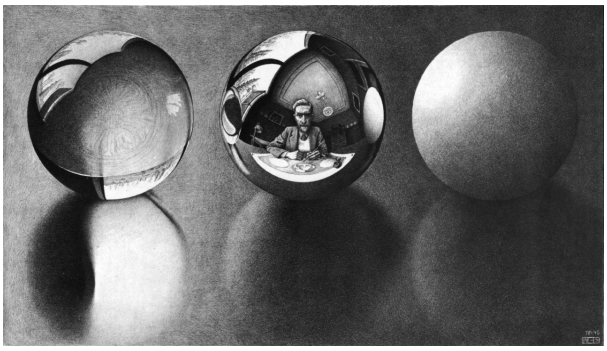


Figure 2: Three Spheres II, M.C. Escher (1946)

- Frequently Asked Questions @GitHub
- First place to go to with a general question
- FAQ is regularly updated with new content

Week 1: introduction to the course

- [X] Review: Prerequisites (see my email on July 18)
- [X] Review: Entry test
- [X] Discussion: Mutual introductions
- [X] Lecture: Course overview
- [X] Practice: course infrastructure
- [X] Group work: Data Science At The End Of Modern Time
- [X] Homework: Assignments
- [ ] Glossary

Prerequisites

Figure 3: SPAM classic

In DSC 482/MTH 445 (Applied maths for data science), I do NOT assume familiarity with R. We will learn the language as we go along. This will necessarily slow us down but this is a math/computing course, so concepts (in this case probability/statistics) are as important as coding. Textbook: Book of R, Part III (ch. 13-16), by T D Davies (NoStarch, 2016).

- General preparation: see the FAQ - R, Emacs + ESS + Org-mode

Bill Gates committed what is called a statistical "Type II" error, when in 2004, he said "Two years from now, spam will be solved." Not rejecting this null hypothesis by Microsoft, arguably lead to underestimating spam (2014: above 70% of all emails sent, 2021: above 45% of all emails sent).

Quick entry test analysis

```
results <- c(10.67, 9.67, 13.83, 7.83, 7.17, 11.67, 11.67, 10.83, 9.67, 11.67, 14.42)
hist_results <- hist(results)
```

```
results <- c(10.67, 9.67, 13.83, 7.83, 7.17, 11.67, 11.67, 10.83, 9.67, 11.67, 14.42)
plot(density(results))
```

Practice - course infrastructure

Useful: take notes! Practice leads to mastery and the practice exercises will often come back to haunt you in the tests.

1. Open a browser
2. Find the GitHub repos (birkenkrahe/dviz and /org)
3. Open the command line terminal
4. Open/close R
5. Open Emacs
6. Find the Emacs tutorial
7. Open/close R inside Emacs
8. Run R in an Org-mode file
9. Close Emacs
10. Close the command line terminal

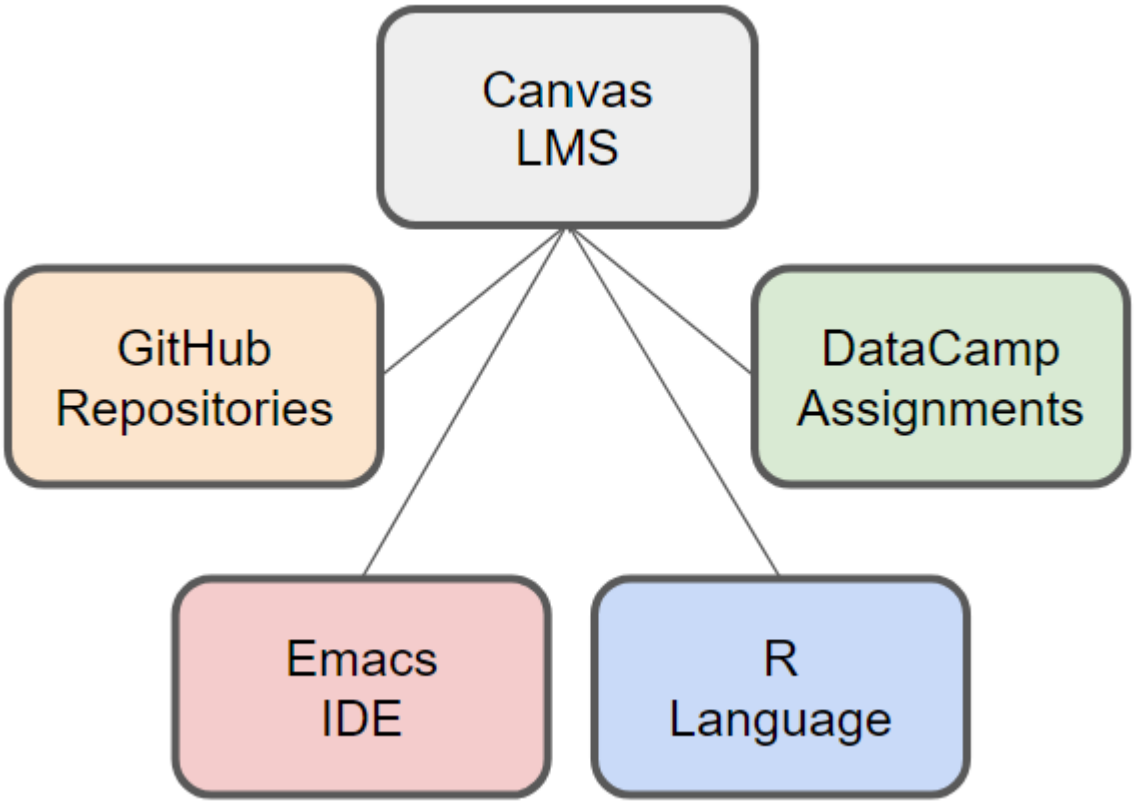
Note: Class room practice completion = 10 points each for active participation.

(Link to practice file in GitHub)

Week 2: getting started

- [X] Quiz 1: course infrastructure
- [X] About: home assignments
- [X] Getting started with projects
- [X] Group work: Data Science At The End Of Modern Time (GitHub)
- [X] Lecture: Data Science At The End Of Modern Time
- [ ] Practice: Running R in an Emacs Org-mode file (GitHub)
- [ ] Lectures: Describing raw data with statistical variables
- [ ] Practice: Raw data and statistical variables (GitHub)
- [ ] Home assignment: summary statistics (DataCamp)

Home assignments - how they work



- Assignment must be completed on time on DataCamp
- Assignment is posted on Canvas (includes the link)
- You loose 1 point for every day of late submission
- Canvas Gradebook is updated manually (with some delay)

Getting started with projects

- Course has 14 participants!
- You'll have to do the project in a team - 2 to 3 people
- We can only accommodate at most 8 projects (last week of term)
- Put your team/ideas into this table (Canvas) by Thursday
- [X] **Who has not yet found a team?**
- [X] **Who has a team but no idea what to do?**
- Reminder: plenty of project opportunities (overview / GitHub issues)

Featured example

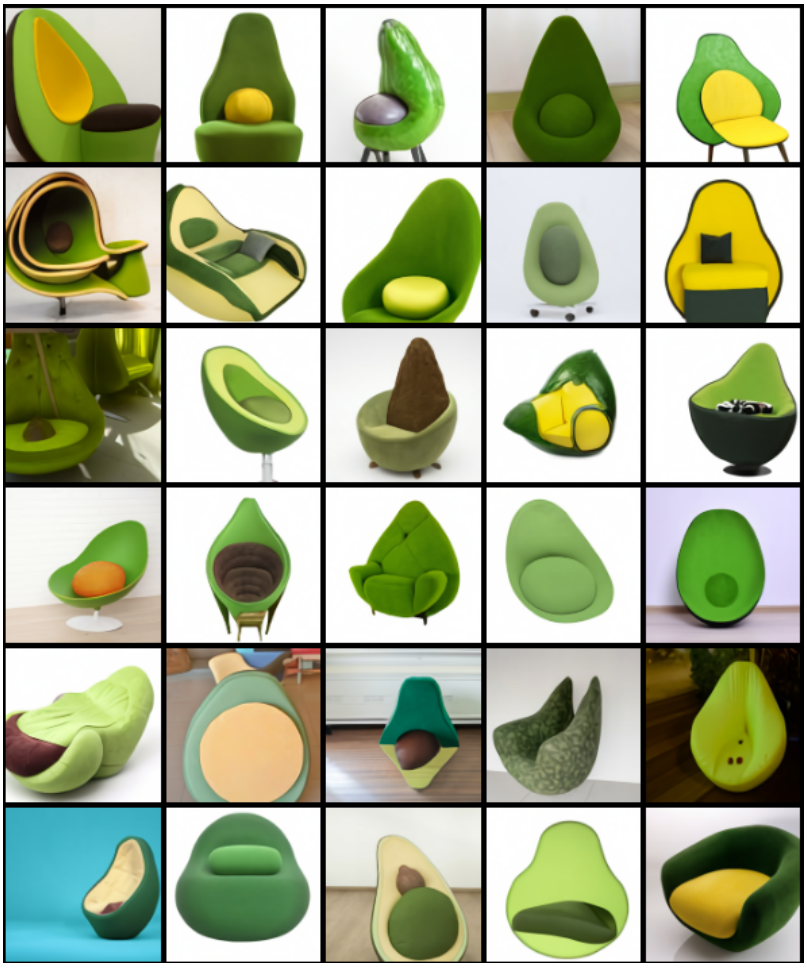


Figure 5: text prompt = an armchair in the shape of an avocado.

- GitHub issue: DALL-E math
- Source: OpenAI - creating images from text)
- DALL-E is a so-called transformer language model (explanation)
- Your project could consist in trying to understand what it is about, place it in context, perhaps clarify some of the math, and relate this to the class
- Avocado example and others

Group exercise: orientation

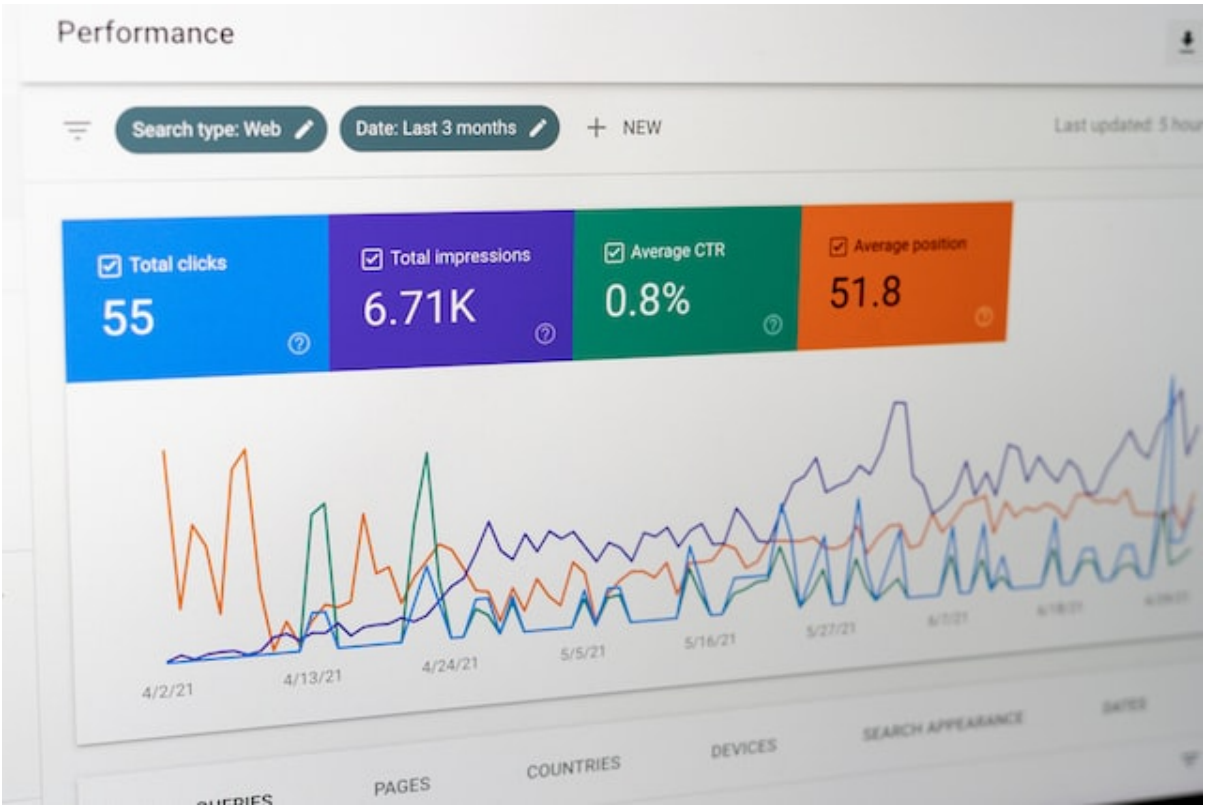
> "Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means." —Bertrand Russell, 1929 Lecture (cited in Bell 1945, 587)

(Results - PDF)

1st sprint review - Wed 1-Sep

- Canvas assignment with submission
- Complete projects overview table in Canvas **today**!
- If you are in > 1 course, you can use the same project idea!

Week 3: elementary statistics



- [X] Due: Quiz 2: week 2
- [X] Reminder: 1st sprint review due September 1st
- [X] Practice: Running R in an Emacs Org-mode file (GitHub)
- [X] Review: DataCamp assignment "Summary Statistics"
- [X] Home assignment: summary statistics (DataCamp)
- [X] Lectures: Describing raw data with statistical variables
- [X] Practice: Raw data and statistical variables (GitHub)

Figure: web page "performance". These are statistics that use summary statistics (e.g. averages) but otherwise they are closely tied to the domain of web traffic monitoring (or SEO - Search Engine Optimization) in order to increase Click-Through-Rate (CTR).

DONE Project: look at Google Analytics

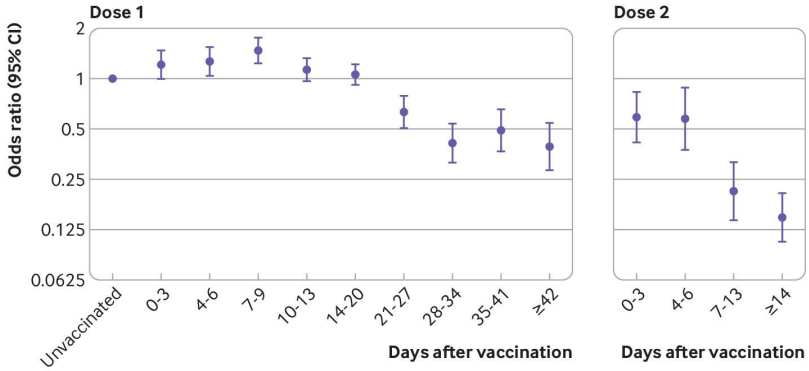
- You won't be able to access analytics.google.com

DONE GNU Emacs: ref cards

- [ ] The power of Dired on one page (v28)
- [ ] The power of Emacs on two pages (v27)

DONE Review: DataCamp "Summary statistics"

- Did you look at the article about COVID-19 vaccines?



- What's a serious limitation of statistics?
- What are "measures of center"? Which ones do you know?
- What are "measures of spread?" Which ones do you know?

- **Limitation of statistics:** cannot be used to find out **why** relationships exist, i.e. does not establish causation
- **Measures of center:** summarize data
  - mean or average
  - median or middle value

    - mode or most frequent value
- **Measures of spread:** indicate variety or clustering
    - range or min/max distance
    - variance or average distance from mean
    - standard deviation or square root of variance

- Next assignment: "probability and distributions"

Figure: "Adjusted odds ratios for confirmed cases of covid-19 by interval after vaccination with Pfizer-BioNTech BNT162b2 before 4 January 2021 in those aged 80 years and older".

"Odds ratios are used to compare the relative odds of the occurrence of the outcome of interest (e.g. disease or disorder), given exposure to the variable of interest (e.g. health characteristic, aspect of medical history). The odds ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome:

- OR=1 Exposure does not affect odds of outcome
- OR>1 Exposure associated with higher odds of outcome
- OR<1 Exposure associated with lower odds of outcome" (Source: nih.gov)

DONE Recap and exercise: data frames

- `[ ]` R functions:
    - `data.frame` - table, column vectors (like SQL)
    - `c` - creating vectors, concatenation
    - `factor` - vectors that hold categorical variables
    - `str` - structure of any R object
    - `$`, `[]` - indexing operators
    - **NEW**: subset

- Test questions:

    - How can you extract a vector named `bar` from a data frame named `foo`? R command: `foo$bar`
    - How can you extract elements with multiple conditions?
    - How can you find out how many rows and columns a data frame has?

    1. `foo$bar` - if you know the column number `N`: `foo[,N]`, e.g. `mtcars[,1]` for the `mpg` column (`N=1`).
    2. By using logical expressions
    3. `dim`, `nrow` x `ncol`, `str`

```
## head(mtcars)
mtcars$mpg
mtcars[,1]
```

```
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4
[16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7
[31] 15.0 21.4
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4
[16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7
[31] 15.0 21.4
```

- `[ ]` Continue completing the practice file

DONE Review: test 2

Match the statistical variable type and the variable.

VARIABLE	TYPE
Weight in lbs.	numeric-continuous
Number of apples on a tree	numeric-discrete
Seniority ("freshman", "junior", "sophomore", "senior")	categorical-ordinal
Employment status ("full-time", "part-time", "unemployed")	categorical-nominal

History of probability and statistics

Match the dominant way of finding out truth, and the historical period.

WORLD-VIEW	PERIOD
Truth is in logic and numbers	Classical period
Truth lies in meditation and in God	Medieval period
Truth is found through experiment	Modern period
Truth is constructed by man	Postmodern period

Data frame value extraction

`df` is a data frame with four variables: `person`, `age` in years, `sex` (`M` or `F`), and `height` in cm. Complete the R command to extract the persons who are taller than 180 cm.

```
df$___ [ df$___ > 180]
```

- [X] `person height`
- [ ] `height person`
- [ ] `persons height`
- [ ] `sex height`

Solution

```
df <- data.frame (
  person = c("Peter", "Lois", "Meg", "Chris", "Stewie"),
  age = c(42, 40, 17, 14, 1),
  sex = factor(c("M", "F", "F", "M", "M")),
  height = c(182, 177, 168, 179, 187))
df
subset(x=df,df$height>180)
df$person[df$height>180]
```

```
  person age sex height
1  Peter  42   M    182
2   Lois  40   F    177
3    Meg  17   F    168
```

```
4  Chris  14   M    179
5 Stewie   1   M    187
  person age sex height
1  Peter  42   M    182
5 Stewie   1   M    187
[1] "Peter"  "Stewie"
```

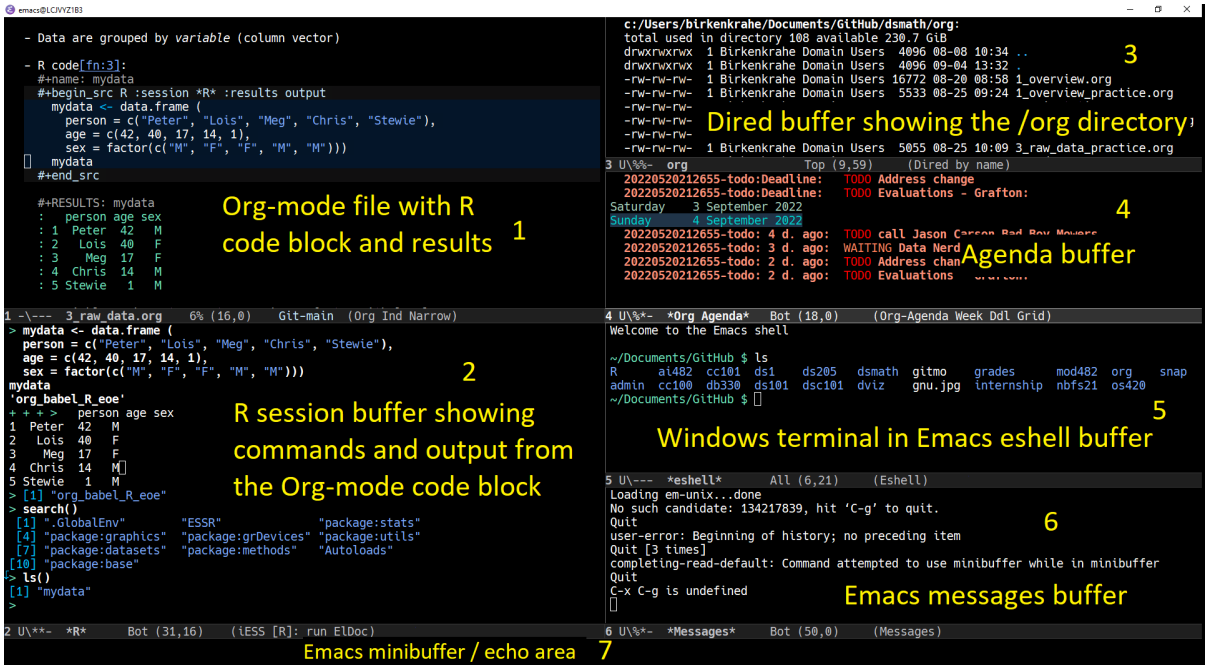
`C-c C-c can do nothing useful here` error

Try `M-x org-mode-restart`.

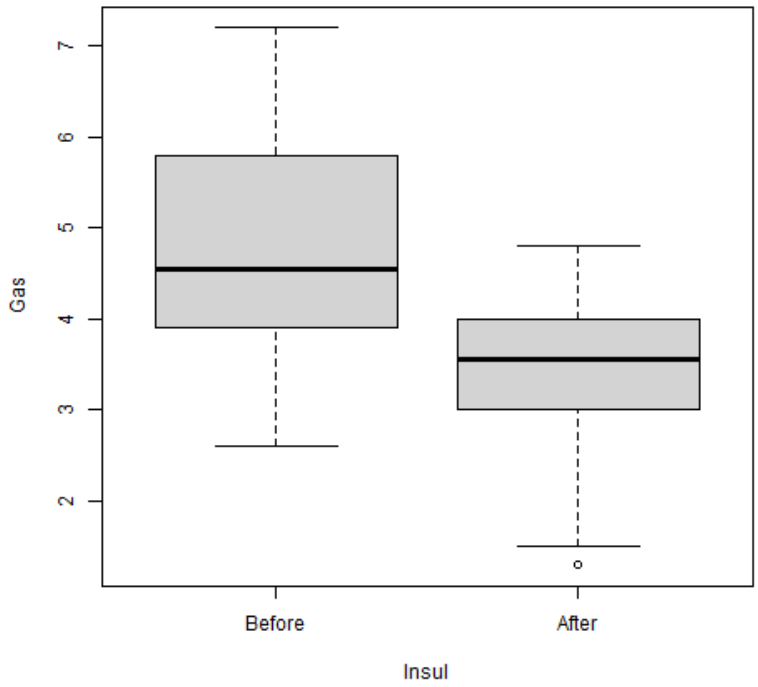
Week 4: describing raw data

- [ ] How Emacs, Org-mode and ESS work together
- [ ] Review test 3 - summary statistics
- [ ] Review 1st sprint review - "pride comes before the fall"
- [ ] Practice: data frames

DONE How Emacs, Org-mode and ESS work together



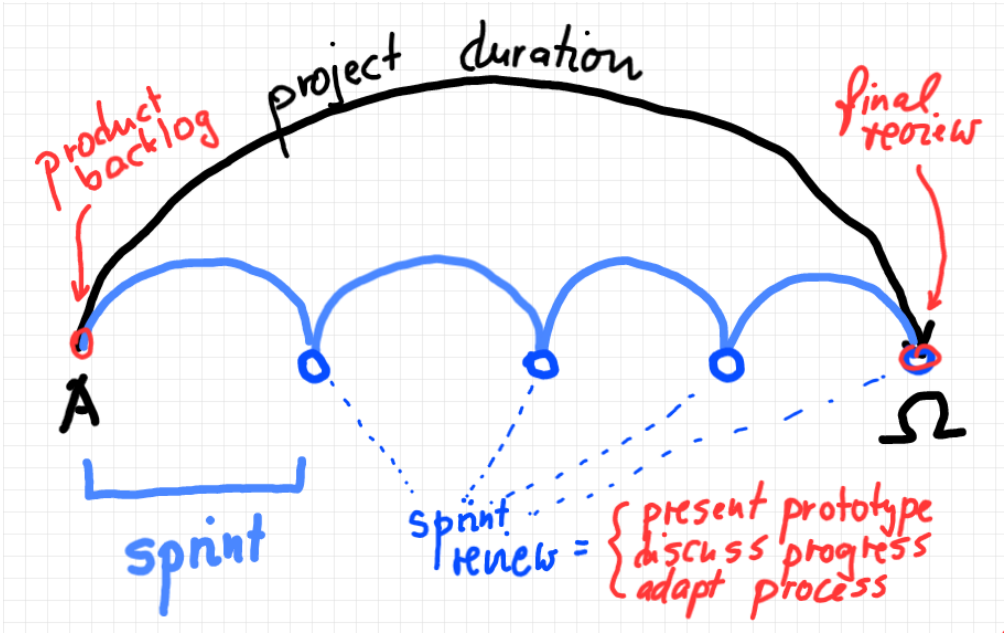
DONE Review test 3 - summary statistics



1. What are descriptive vs. inferential statistics? (83%)
2. What are the limitations of statistics? (67%)
3. Which plots visualize measures of spread? (50%)

On (2): check Judeah Pearl's "Book of Why"

DONE Review: 1st sprint review



"Pride"

- Pride according to the Oxford dictionary:

  »A feeling of being pleased or satisfied that you get when you or people who are connected with you have **done something well** or **own something** that other people **admire**.«

  In other words: if you cannot identify what you're proud of, you either haven't done anything well, or you're not aware of it, which won't do.

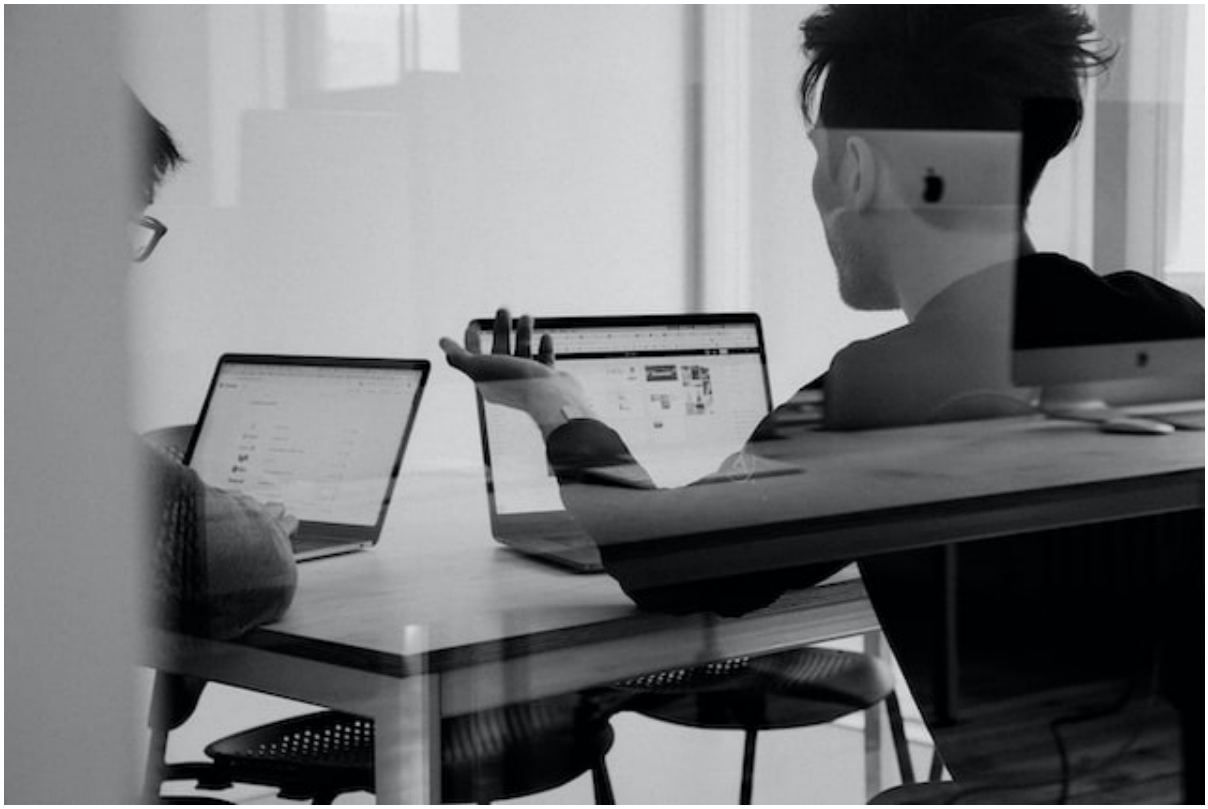
- Of course, *"pride comes before the fall"* (Proverbs 16:18), but in the context of Scrum, it is only one of several qualities to assess the results of a sprint.

"References"

No.	Year	Full Reference	Relevance	Credibility	Possible Use
1	2015	Elston, C. and Morris, N. (2015). Making MOOCs collaboratively: working effectively with stakeholders. In: Proceedings of the European MOOC Stakeholder Summit 2015, pp. 28-31. Mons, Belgium: Universite catholique de Louvain.	Medium (intro)	Medium (conf. Proceed., special conf.)	In introduction: to validate & illustrate the need for a lot of resources when creating MOOCs.
2	2005	Wang, F., Hannafin, M. J. (2005). Design-based research and technology-enhancing learning environments. In: Educational Technology Research and Development, 53(4), pp. 5-23.	High (Method)	High (citations, Unique, scholarly journal, monogr., ranked)	In methods section: central technical reference for the method used for this research.
3	2004	Hevner, A.R., March, S.T., Park,J. And Ram, S. (2004). Design Science in IS Research, MISQuarterly, 28(1), pp. 75-105.	Medium (Method; review article)	High (top-ranked journal in this field)	In methods section: technical reference for the method used in this research; validates the use of this method in the IS field.
4	2009	Birkenkrahe, M., Mundt, M. (2009). From crisis to creativity: undergraduates craft their own online learning modules. In: International Journal for Innovation in Education, 1(1), pp. 96-119.	High (Discussion, own earlier work)	Medium (non-ranked, scholarly journal, own publication)	In introduction & discussion sections: to explore the differences two different experiments.
5	2008	Grzega, J., Schöner, M. (2008). The didactic model LdL (Lernen durch Lehren) as a way of preparing students for communication in a knowledge society. Journal of Education for Teaching: International research and pedagogy, 34(3), pp. 167-175	Medium (Discussion)	Medium (scholarly journal, non-ranked)	In abstract & discussion: basic didactic theoretical approach to anchor the findings. Implicit (non-graphical) model.

- Some of you mentioned references, few provided any
- To do this week: Literature Review with cheat sheet.
- Download it from GitHub, find at least 5 references, label them according to the categories (esp. relevance and credibility), and provide a complete, consistent set of citations.

"Questions"



- You should always use an opportunity to ask the customer/product owner anything, even if it's something simple. (**Why?**)
- Good question are specific, open (not closed as in yes/no), and use the qualities (as in: variables!) that you're after, e.g. "What do you like about me in terms of punctuality, systematic work, appearance…"
- A question is specific if you can immediately use it to take an action!

Only one team asked questions at all (Nikkolette/Wyatt):

- *What was the hardest part so far for you?*
- *What was the most interesting part you have found/want to find?*

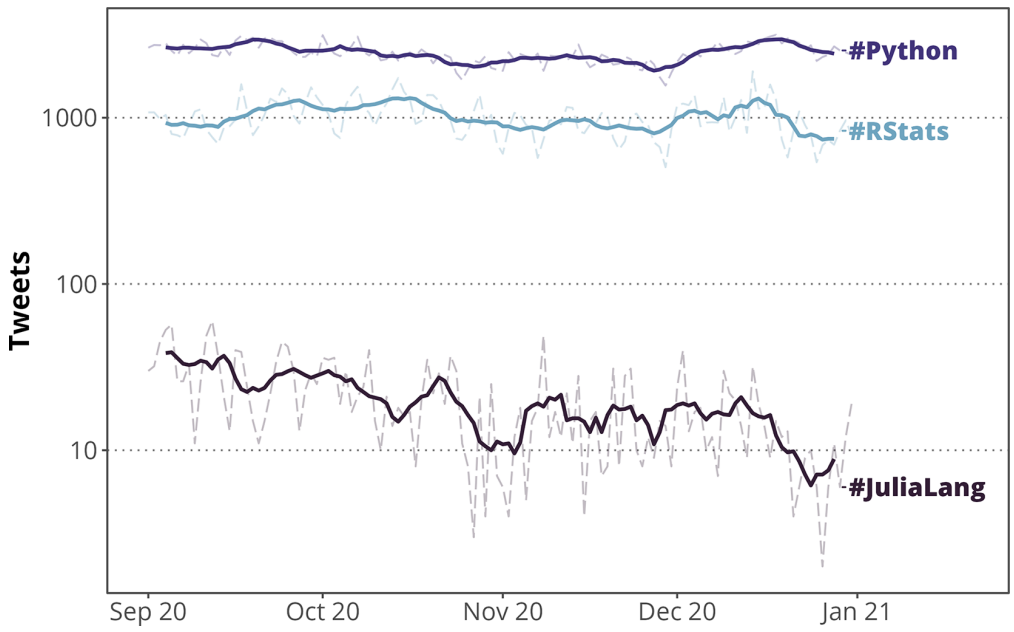
Better next time!

1. Deliver more than the bare minimum **generously**
2. Try to make your project great by working **systematically**
3. If you have a team, split up the work **meaningfully**
4. If you have any questions, ask others and me **bravely**
5. Complete the (optional) literature review **diligently**

DONE Practice: raw data stats (30 min)

- Go to the practice file (GitHub: tinyurl.com/23f9uz8s)
- Complete the practice exercise on **data frames**
- You can find example code in the lecture (GitHub: tinyurl.com/2am222mh)

NEXT **Ten simple rules for teaching yourself R (Lawlor et al, 2022)**



- Written for biologists, not computer scientists. Relevant community: bio and health science stats (Prof Chapman sent me this yesterday)
- I support some but not all recommendations:
    1. "Build skills with low-pressure projects" (i.e. play around)
    2. Don't worry about style but worry about documentation
    3. "Join the R community" - I also use Twitter
    4. "Read others' code, and share yours" - use GitHub
    5. "Don't box yourself in" - use languages for what they're good at

Week 5: summary statistics

Figure 19: Charles II of England (1630-1685)

- [x] 1654: letters between Blaise Pascal and Pierre de Fermat
- [x] Featured application: retraction watch
- [x] Off-topic: Laporta algorithm (Feynman diagram evaluation)
- [x] Review: test 4
- [x] Review: DataCamp lesson probability and distributions
- [x] Review: logical flag vectors
- [x] Practice: statistical variables (continued)
- [x] Lecture/practice: summary statistics

Review: test 4 - raw data, probability and stats

- [x] Longitude/latitude are what kind of data?
- [x] When researching, do you always need a "literature review"?

  - "Literature review" as a type of paper is the most useful paper you can find as a beginner - look for one in your project area!
- [X] Which activities connect "population" and "sample"?

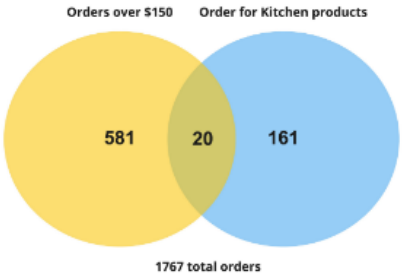
Review: probability and distributions (DataCamp)

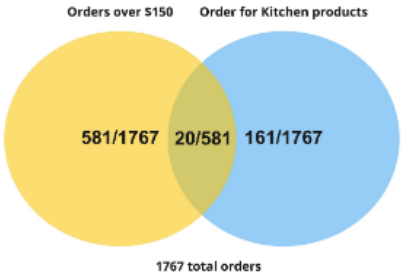
1. What is the conditional probability for an event B given that an event A has already happened (as a formula)

   Formula: P(B|A) = P(A and B) / P(A)

2. How can you visualize the conditional probability formula for events A and B?

   Example: A = Order for kitchen products, B = Orders over $150





$$P(Kitchen|Order > 150) = \frac{\frac{20}{1767}}{\frac{581}{1767}}$$

$$P(Kitchen|Order > 150) = \frac{20}{581}$$

3. A men's soccer team plays soccer zero, one, or two days a week:

   - the probability that they play zero days is .2,
   - the probability that they play one day is .5, and
   - the probability that they play two days is .3.

     What is the long-term average or expected value, μ, of the number of days per week that the men's soccer team plays soccer?

x = DAYS	P(x)
0	0.2
1	0.5
2	0.3

Expected value: E(DAYS) = μ = ∑ x P(x) = 0 * 0.2 + 1 * .5 + 2 * .3 = 1.1

```
## number of days the team plays per week
x <- c(0,1,2)  # events
p_x <- c(0.2, 0.5, 0.3) # probability per event
```

```
mu <- sum(x * p_x) # expected value
paste("expected value: ", mu)
```

4. What is the *law of large numbers*?

   As the size of your sample increases, the sample mean will approach the expected value (the population average).

```
x <- sample(rep(1:6),size=10,replace=TRUE)
hist(x, xlab="10 rolls, fair dice", main="die roll")
abline(v = mean(x), col="red",lwd=2)
abline(v = sum(x/6),col="blue",lwd=2)
```

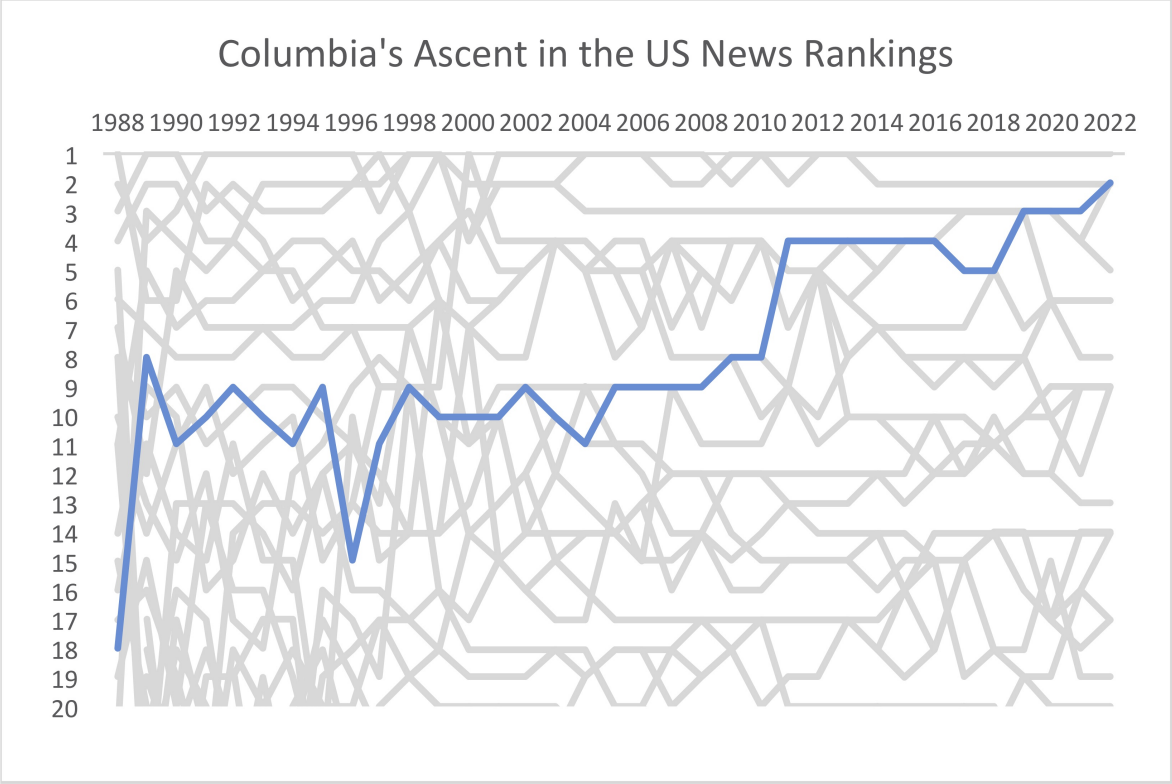
5. What is the probability that a baby will be born between midnight and 8 am? (If all hours are equally probable.)

   A day has 24 hours - midnight to 8 am is 8/24 or 1/3, so 33%.

Raw data: statistical variables (practice)

Featured: university ranking (issue)

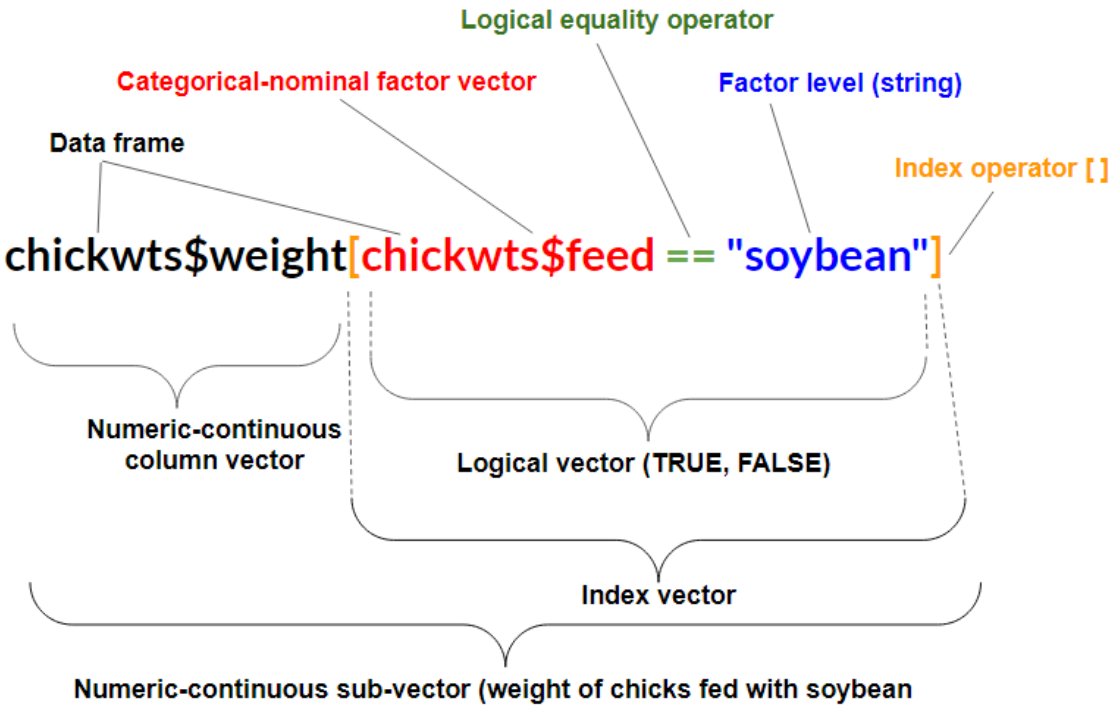
- Columbia U math professor uncovers stats lies
- Columbia U moved up from 18th to 2nd between 1988 and 2022



Review: logical flag vectors

Can you name and explain the 9 elements of this expression?

```
chickwts$weight[chickwts$feed == "soybean"]
```



R code - logical flag vector

```
str(chickwts) # structure of the chickwts data set
```

```
chickwts$weight # display numerical column vector
```

```
chickwts$feed # display categorical-nominal factor vector
```

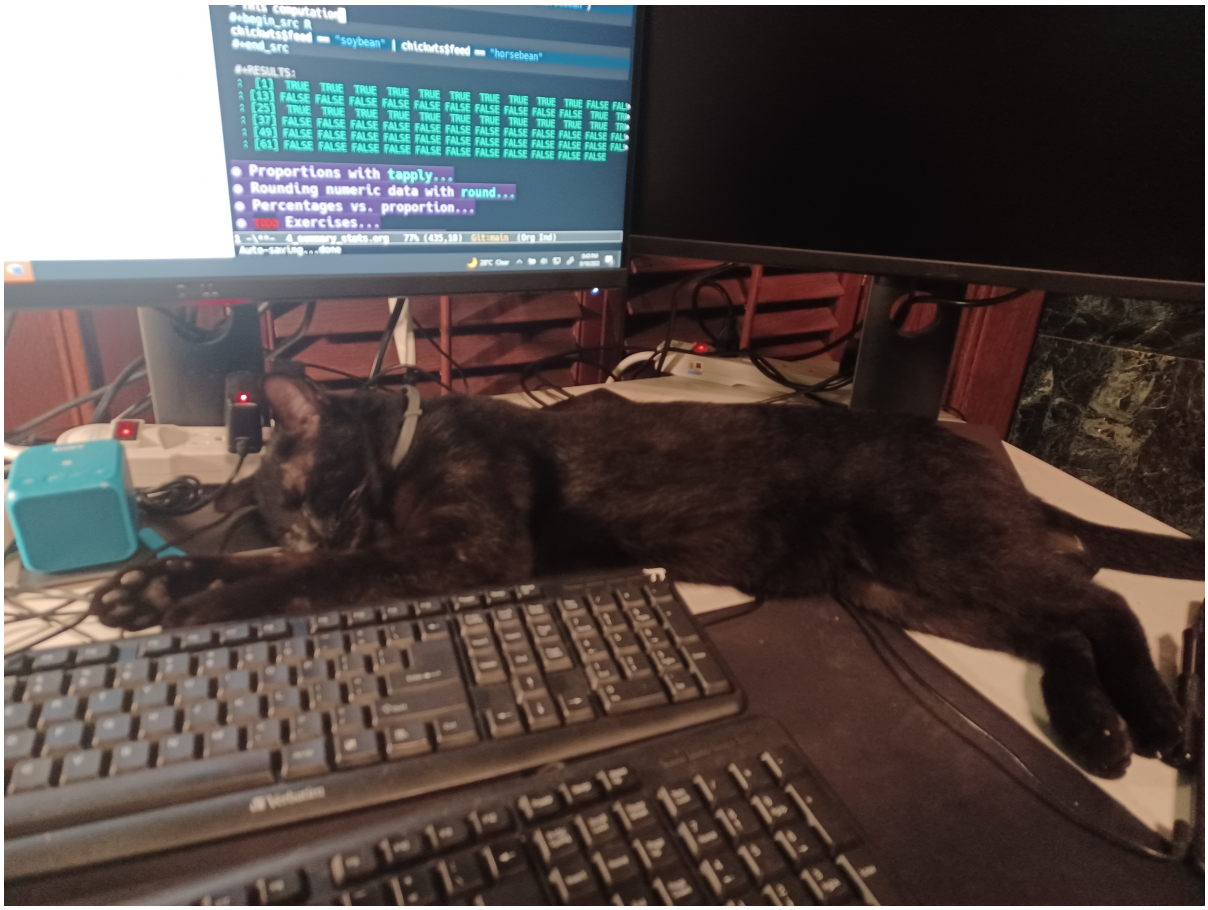
```
chickwts$feed == "soybean" # display the "soybean" level of feed
```

```
chickwts$weight[chickwts$feed == "soybean"] # show weight of chicks fed on soybean
```

```
which(chickwts$feed == "soybean")  # get index values for chicks fed on soybean
chickwts$weight[which(chickwts$feed == "soybean")] # show weight of chicks fed on soybean
```

```
str(chickwts) # data frame structure
chickwts$feed # factor vector, categorical-nominal
chickwts$feed == "soybean" # logical vector
which(chickwts$feed == "soybean") # numveric index vector
chickwts$weight[chickwts$feed == "soybean"] # numeric vector
```

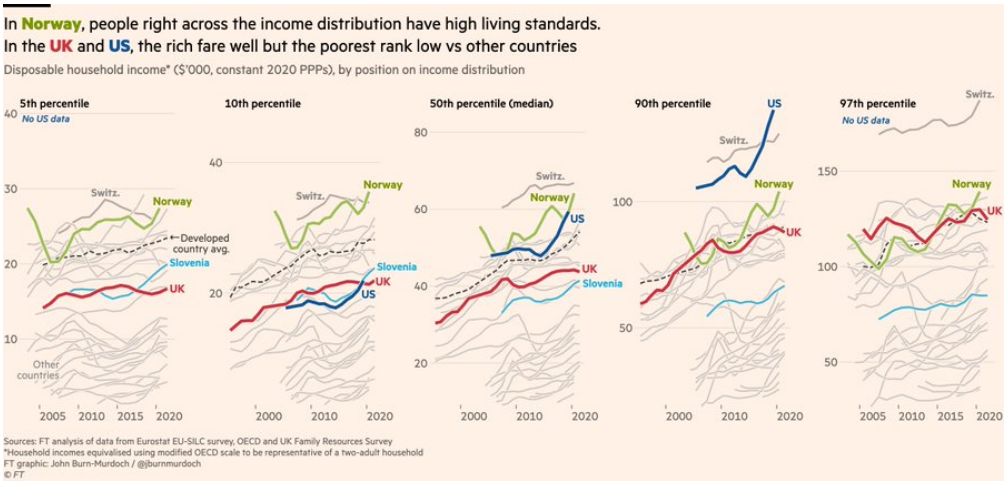
Week 6: counts, proportions, percentages



- [ ] Featured applications
- [ ] DataCamp deadline extended once more (23 Sept 11:59pm)
- [ ] Lecture/practice on summary statistics (continued)
- [ ] Home assignment until Thursday (Org-mode file)

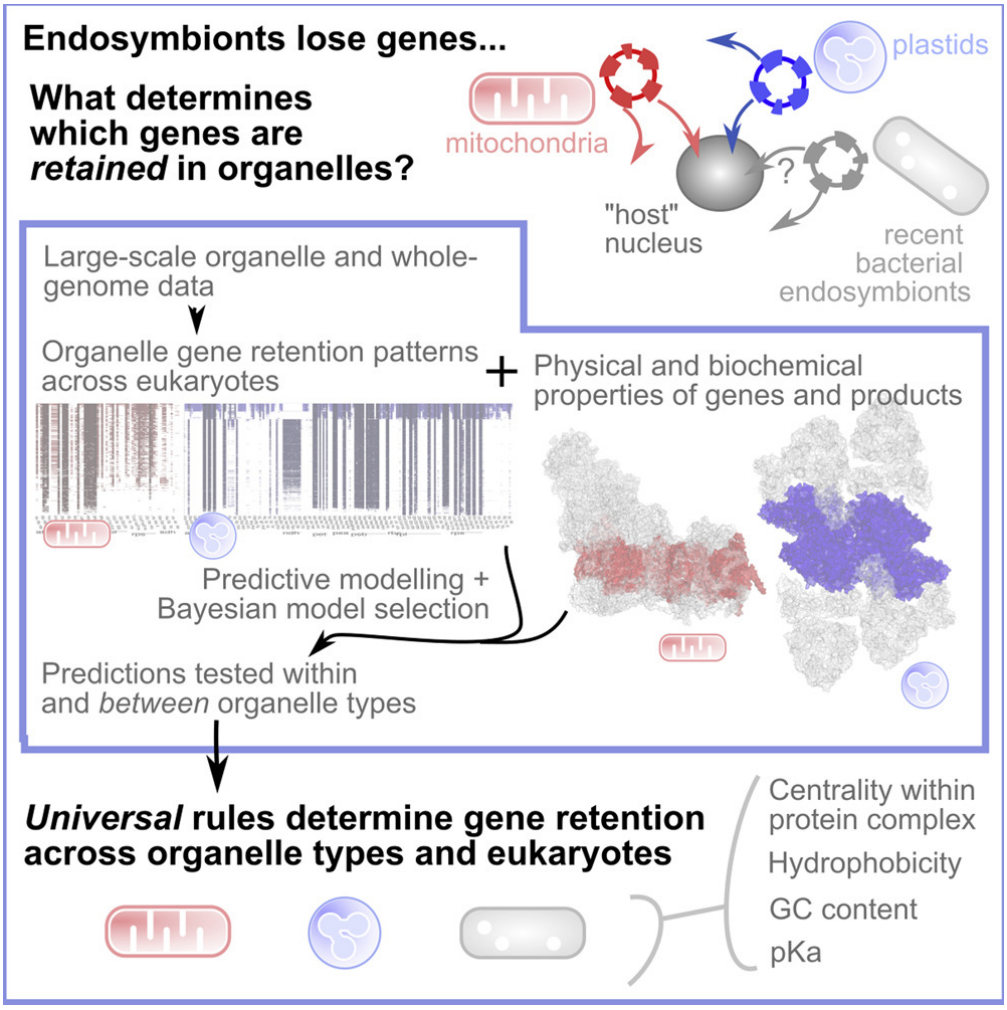
Featured applications: (issues)

- Investigation of income equality in US and UK using percentiles



- Data science reveals universal rules shaping cells' power stations

  "The scientists took a data-driven approach. They gathered data on all the organelle DNA that has been sequenced across life. They then used modeling, biochemistry, and structural biology to represent a wide range of different hypotheses about gene retention as a set of numbers associated with each gene. Using tools from data science and statistics, they asked which ideas could best explain the patterns of retained genes in the data they had compiled—testing the results with unseen data to check their power."



TODO Lecture/practice: summary statistics (cont'd)

- Open your Emacs Org-mode practice file `stats.org`
- At the top, below the `#+PROPERTY:` line, add the line:

  ```
  #+STARTUP: overview hideblocks indent inlineimages
  ```

- Now, in the body of the document, add headlines like this:

  ```
  * Getting started
  ** Getting bored
  ```

- Go to the bottom of your file with `M->`
- Add another headline for the next section:

  ```
  * Category subsets with ~tapply~
  ```

- Additional code blocks should go below this headline

TODO Home assignment: summary statistics exercises


(Image: celebrate the German garden gnome!)

TODO DC Review: Central Limit Theorem (23 Sept)

References

Lawlor J, Banville F, Forero-Muñoz N-R, Hébert K, Martínez-Lanfranco JA, Rogy P, et al. (2022) Ten simple rules for teaching yourself R. PLoS Comput Biol 18(9): e1010372. https://doi.org/10.1371/journal.pcbi.1010372

Author: Marcus Marcus Speh Birkenkrahe
Created: 2022-09-22 Thu 18:01