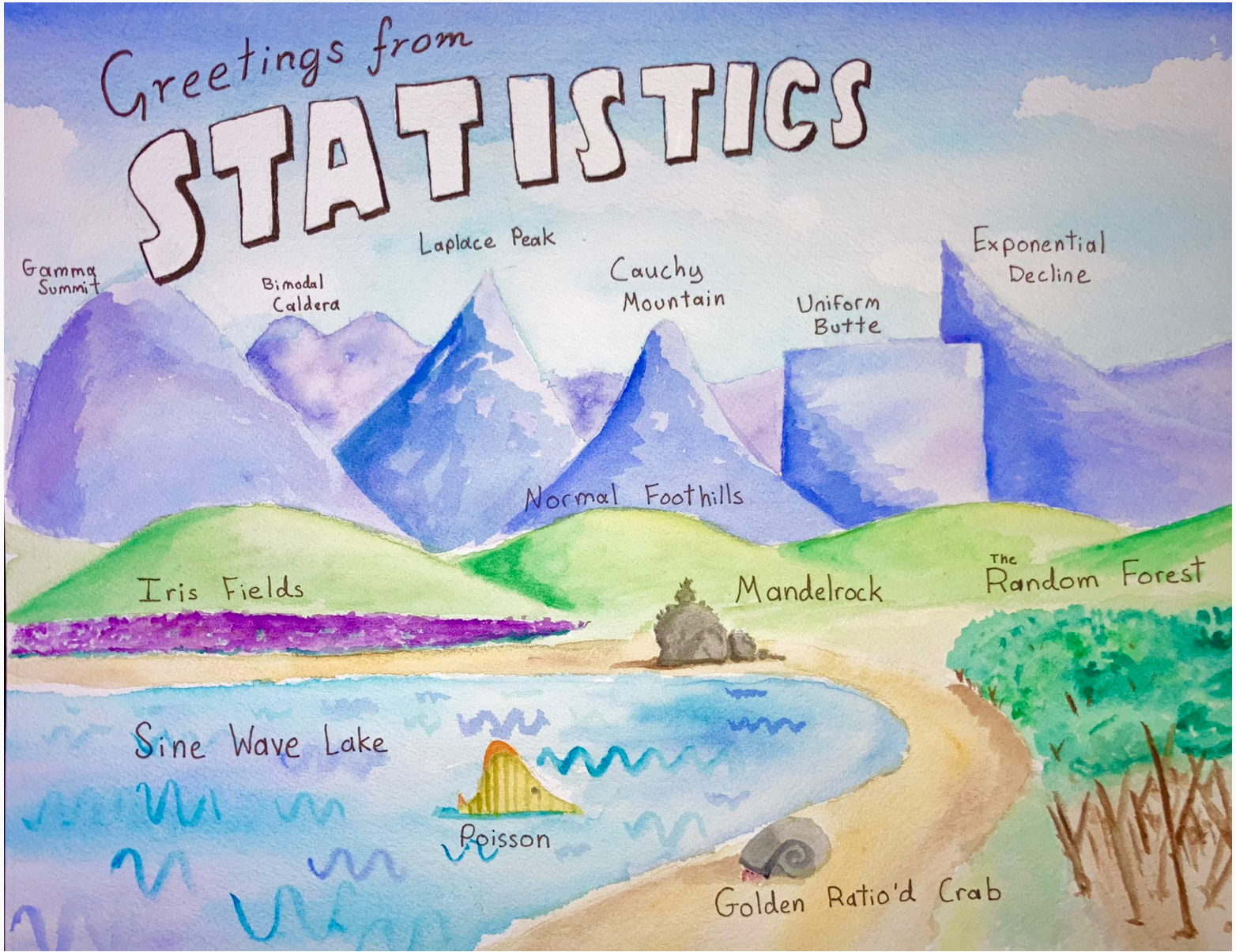
Introduction to DATA 606

Statistics & Probability for Data Analytics

Jason Bryer, Ph.D. and Angela Lui, Ph.D.

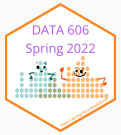
Spring 2022


Greetings from
STATISTICS
Gamma Summit
Bimodal Caldera
Laplace Peak
Cauchy Mountain
Uniform Butte
Exponential Decline
Normal Foothills
Iris Fields
Mandelrock
The Random Forest
Sine Wave Lake
Poisson
Golden Ratio'd Crab

@skyetetra

Agenda

- About your instructors
- Syllabus
- Class meetups
- Course Schedule
- Assignments (how you will be graded)
  - Participation
  - Labs
  - Data Project
  - Exams
- Software
  - The `DATA606` R Package
  - Using R Markdown



A little about Jason...

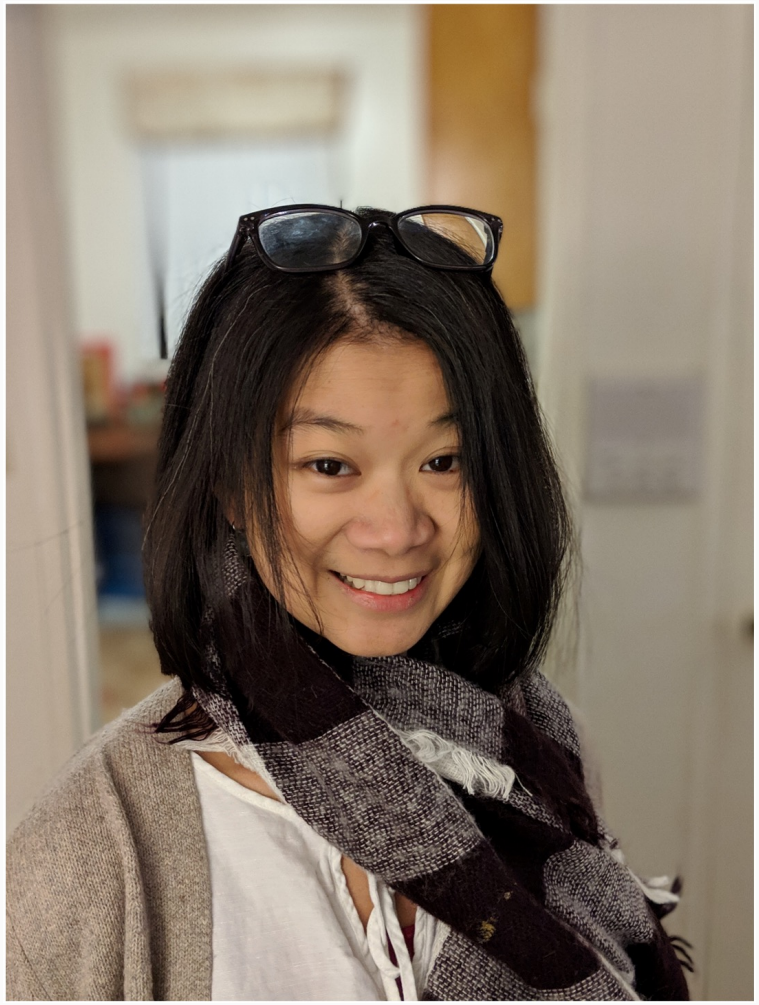
- Assistant Professor at CUNY in Data Science and Information Systems
- Principal Investigator for a Department of Education Grant (part of their FIPSE First in the World program) to develop a Diagnostic Assessment and Achievement of College Skills (www.DAACS.net)
- Authored over a dozen R packages including:
  - likert
  - sqlutils
  - timeline
- Specialize in propensity score methods. Three new methods/R packages developed include:
  - multilevelPSA
  - TriMatch
  - PSAboot
- Developer of a data dashboard for the NYS Office of Special Education and TAP for Data at Cornell University: https://data.osepartnership.org

Also a Father...

Runner...

And photographer.

A little about Angela...



NYU

HUNTER

UNIVERSITY AT ALBANY
State University of New York

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

CUNY School of Professional Studies | CUNY

DAACS
Diagnostic Assessment & Achievement of College Skills

DATA 606
Spring 2022

Teaching Experience

- Introduction to Statistics in Social Sciences
- Special Issues in Testing
- Evaluation
- Motivation in Education
- Introduction to the Psychological Processing of Schooling
- Educational Psychology in Adolescent Development

Homeowner

Syllabus

Syllabus and course materials are here: https://spring2022.data606.net

The site is built using the Blogdown R package and hosted on Github. Each page of the site has a "Improve this page" link at the bottom right, use that to start a pull request on Github.

We will use Blackboard primary for submitting assignments only. Please submit:

- A PDF or link to the built HTML (e.g. Rpubs, Github)

PDFs are preferred for the homework as there is some LaTeX formatting in the R markdown files. The `tinytex` R package helps with install LaTeX, but you can also install LaTeX using MiKTeX (for Windows) and BasicTeX (for Mac) See this page for more information: https://spring2022.data606.net/course-overview/software/

Meetups

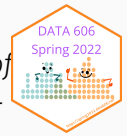
We will have meetups on Wednesday evenings at 8:00pm.

Meetups will be recorded and made available the next day on the course website.

Though attending live is not strictly required, **We expect everyone to watch the lectures during the week.** I use the class meetups to convey important information and announcements. Very often I will cover some topics not in the textbook. Students who attend the meetups tend to do well on the assignments.

One Minute Papers - Complete the one minute paper after each Meetup (whether you watch live or watch the recordings). It should take approximately one to two minutes to complete. This allows me to 1) verify you have attended/watch the meetup and 2) get feedback about what you learned and what you may still be unclear.
Link: https://forms.gle/cdit7TEfNTdJyozP6

Please note: *Students who participate in this class with their camera on or use a profile image are agreeing to have their video or image recorded solely for the purpose of creating a record for students enrolled in the class to refer to, including those enrolled students who are unable to attend live. If you are unwilling to consent to have your profile or video image recorded, be sure to keep your camera off and do not use a profile image. Likewise, students who un-mute during class and participate orally are*



Schedule

Start	End	Topic
Friday, January 28, 2022	Sunday, February 06, 2022	Chapter 1 - Intro to Data, R, and RStudio
Monday, February 07, 2022	Sunday, February 13, 2022	Chatper 2 - Summarizing Data
Monday, February 14, 2022	Sunday, February 20, 2022	Chapter 3 - Probability
Monday, February 21, 2022	Sunday, February 27, 2022	Chapter 4 - Distributions
Monday, February 28, 2022	Sunday, March 13, 2022	Chatper 5 - Foundation for Inference
Wednesday, March 09, 2022	Tuesday, March 15, 2022	Midterm
Monday, March 14, 2022	Sunday, March 20, 2022	Chapter 6 - Inference for Categorical Data
Monday, March 21, 2022	Sunday, March 27, 2022	Chapter 7 - Inference for Numerical Data
Monday, March 28, 2022	Sunday, April 10, 2022	Chapter 8 - Linear Regression
Monday, April 11, 2022	Sunday, May 01, 2022	Chapter 9 - Multiple & Logistic Regression
Monday, May 02, 2022	Wednesday, May 11, 2022	Intro to Bayesian Analysis
Wednesday, May 11, 2022	Sunday, May 15, 2022	Final Exam

Textbooks

Diez, D.M., Barr, C.D., & Çetinkaya-Rundel, M. (2019). *OpenIntro Statistics (4th Ed)*.

This will be our primary textbook for most of the semesters. Our goal is to cover all the chapters.

Navarro, D. (2018, version 0.6). *Learning Statistics with R*

This textbooks has a chapter on Bayesian analysis that we will use at the end of the semester.

Assignments

- Participation (10%)
  - DAACS
  - One Minute Papers
- Labs (35%)
  - Labs are designed to introduce to you doing statistics with R.
  - Answer the questions in the main text as well as the "On Your Own" section.
- Data Project (30%)
  - This allows you to analyze a dataset of your choosing. Projects will be shared with the class. This provides an opportunity for everyone to see different approaches to analyzing different datasets.
- Exams
  - Midterm (10%)
  - Final exam (15%)

Communication

- Slack Channel: https://data606spring2022.slack.com
    - Click here to join the group
- There is a general CUNY MSDS Slack channel click here to join it.
- Github Issues - Use this for issues or problems with the `DATA606` package: https://github.com/jbryer/DATA606/issues
- Email: jason.bryer@cuny.edu and angela.lui@cuny.edu
- Phone/Zoom: Please email to schedule a time to meet.
- Office hours by appointment.

Software

This is an applied statistics course so we will make extensive use of the R statistical programming language. You have two options for using R in this course:

- CUNY SPS has an RStudio Server that you can access using a browser: https://rstudio.sps.cuny.edu
  You will use your CUNY login credentials to log in.

- Install R and RStudio on your own computer. I encourage everyone to do this at some point by the end of the semester. I have instructions on the course website here: https://spring2022.data606.net/course-overview/software/

You will also need to have LaTeX installed as well in order to create PDFs. The `tinytex` R package helps with this process:

```
install.packages('tinytex')
tinytex::install_tinytex()
```

DATA 606 Package

The `DATA606` R package contains many data sets and functions we will use throughout the semester. It also has a `startLab` function that will copy each of the labs to your current working directory. Use the following commands to install the package (only necessary once per R installation):

```
remotes::install_github('jbryer/DATA606')
```

To start the first lab...

```
DATA606::startLab('Lab1')
```

This will copy the R markdown file and any supporting files to your current working directory. Use the "Knit" button in R Studio to build a PDF of the document.

Next steps...

Before Monday (January 31st):

- Complete this Google form: https://forms.gle/ikB64cJjcGUadqZWA
- Create an account at https://my.daacs.net and complete the self-regulated learning assessment
- Join the Slack channel

Then:

- Start Lab 1 (due February 6th)

Good luck with the semester!

jason.bryer@cuny.edu
angela.lui@cuny.edu
data606spring2022.slack.com
@jbryer
@angelalui11
@jbryer
spring2022.data606.net