Introduction to DATA 606

Statistics & Probability for Data Analytics

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

Fall 2024

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 to develop and test the Diagnostic Assessment and Achievement of College Skills (www.DAACS.net)
- Authored over a dozen R packages including:
    - likert
    - ShinyQDA
    - DTedit
    - login
- Specialize in propensity score methods. Three new methods/R packages developed include:
    - multilevelPSA
    - TriMatch
    - PSAboot

Also a Father...

DATA 606 Fall 2024

Runner...

DOPEY CHALLENGE 2020
20602
JASON

2019
TCS NEW YORK CITY MARATHON
NEW YORK ROAD

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 Fall 2024

Teaching Experience

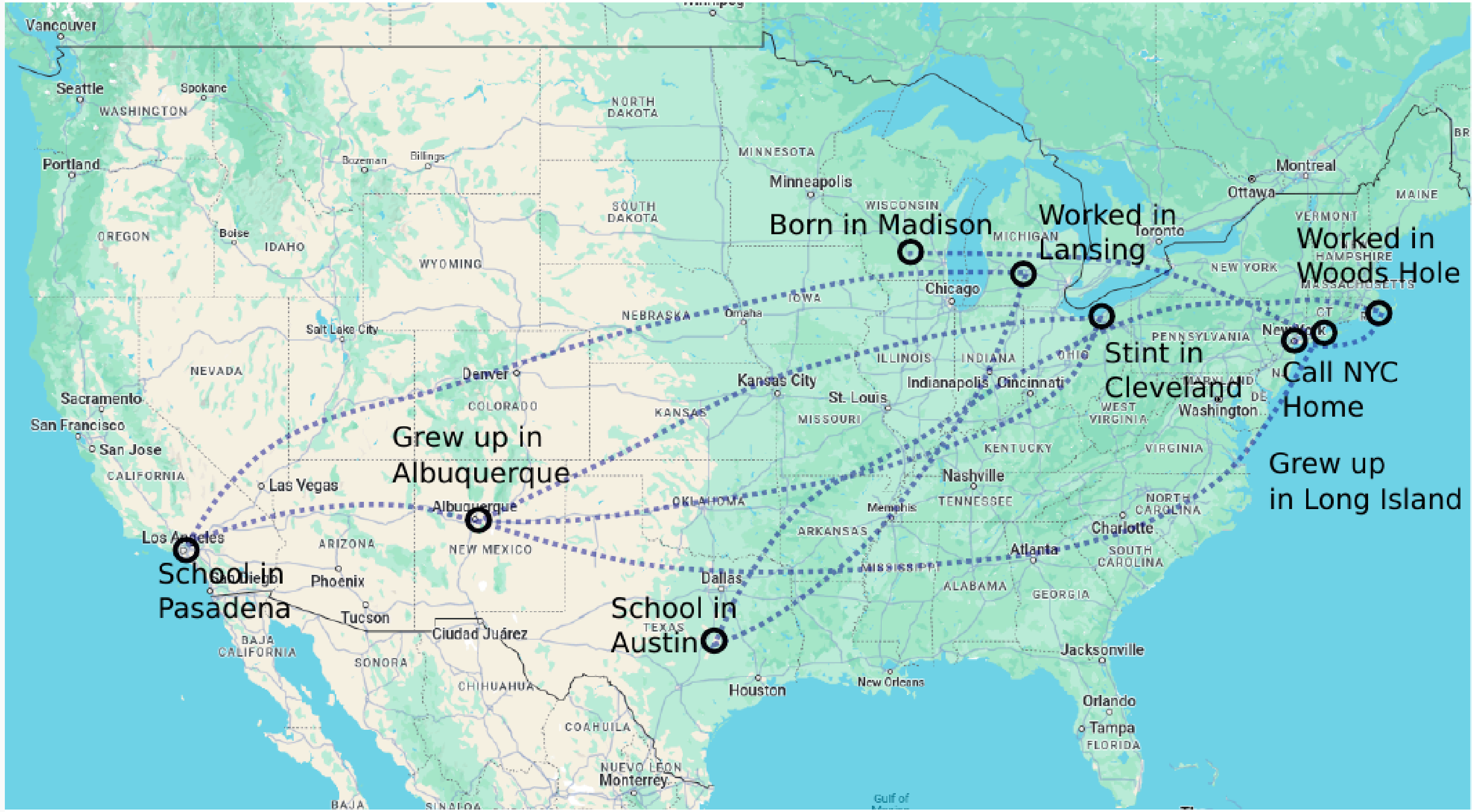
- Introduction to Statistics in Social Sciences
- Special Issues in Testing
- Evaluation
- Motivation in Education
- Introduction to the Psychological Processing of Schooling
- Educational Psychology in Adolescent Development

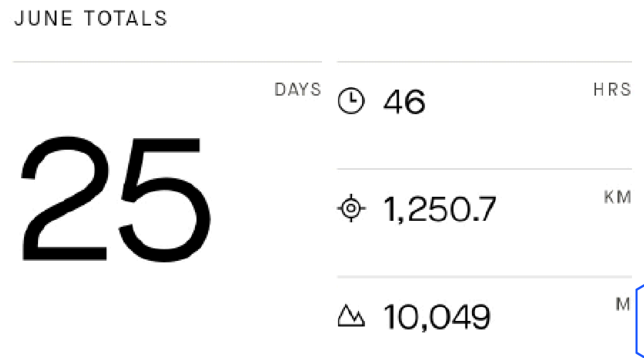
Homeowner

A little about George….

- Doctoral Lecturer at CUNY SPS
- Past research experience:
  - Plasma physics and Applied Mathematics (NYU)
  - Ecology and Evolutionary Biology (Princeton)
- Have expertise and maintain active research interests in using Bayesian methods and genomic data to improve global scale ocean and climate models
- Taught math at NYU

Lived all over, loves cycling





JUNE TOTALS

DAYS

25

46 HRS

1,250.7 KM

10,049 M

Syllabus

Syllabus and course materials are here: https://fall2024.data606.net

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

We will use Brightspace 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 `tineytex` 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://fall2024.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.

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 agreeing to have their voices recorded. If you are not willing to consent to have your voice recorded during class, you will need to keep your mute button activated and communicate exclusively using the "chat" feature, which allows students to type questions and comments live.*

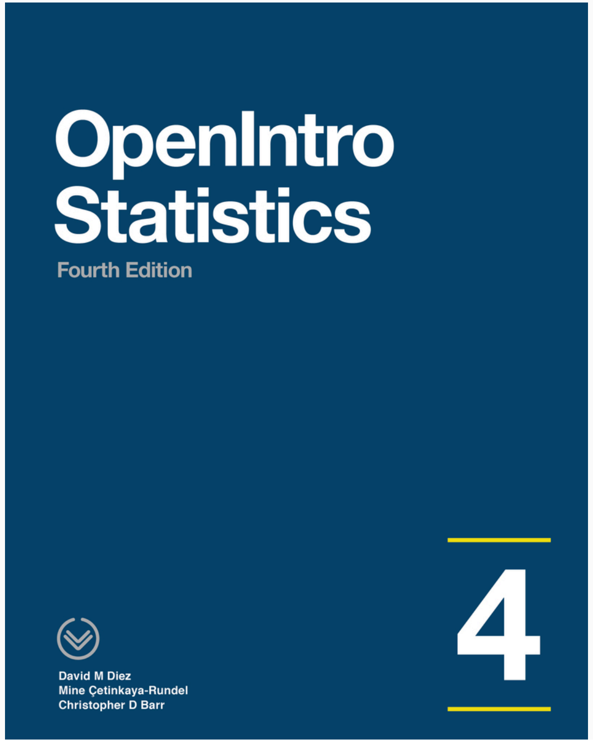
Schedule

Start	End	Topic
Wednesday, September 04, 2024	Tuesday, September 10, 2024	Chapter 1 - Intro to Data, R, and RStudio
Wednesday, September 11, 2024	Tuesday, September 24, 2024	Chatper 2 - Summarizing Data
Wednesday, September 25, 2024	Tuesday, October 01, 2024	Chapter 3 - Probability
Wednesday, October 02, 2024	Tuesday, October 08, 2024	Chapter 4 - Distributions
Wednesday, October 09, 2024	Tuesday, October 15, 2024	Chatper 5 - Foundation for Inference
Wednesday, October 16, 2024	Sunday, October 20, 2024	Midterm
Wednesday, October 16, 2024	Tuesday, October 22, 2024	Chapter 6 - Inference for Categorical Data
Wednesday, October 23, 2024	Tuesday, October 29, 2024	Chapter 7 - Inference for Numerical Data
Wednesday, October 30, 2024	Tuesday, November 05, 2024	Chapter 8 - Linear Regression
Wednesday, November 06, 2024	Tuesday, December 03, 2024	Chapter 9 - Multiple & Logistic Regression
Wednesday, December 04, 2024	Tuesday, December 10, 2024	Intro to Bayesian Analysis
Wednesday, December 11, 2024	Sunday, December 15, 2024	Final Exam

Textbooks

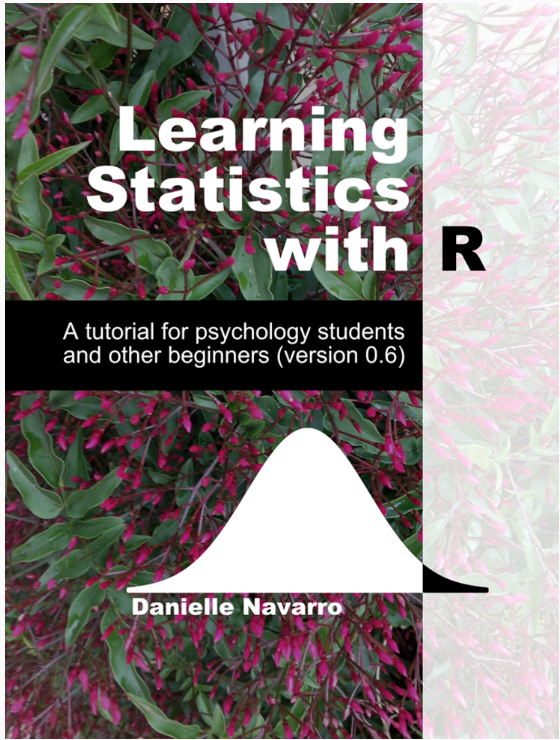
Diez, D.M., Barr, C.D., & Çetinkaya-Rundel, M. (2019). *OpenIntro Statistics (4th Ed).*

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.


Learning Statistics with R
A tutorial for psychology students and other beginners (version 0.6)
Danielle Navarro

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://data606fall2024.slack.com
    - Click here to join the group
- Email: jason.bryer@cuny.edu, angela.lui@cuny.edu, and george.hagstorm@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.

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://spring2024.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:

```r
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):

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

To start the first lab...

```r
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 (September 2nd):

- Complete this Google form: https://forms.gle/zJzxuPLbYtib3Lux5
- Go to https://cuny.daacs.net and complete the self-regulated learning assessment
- Join the Slack channel

Then:

- Start Lab 1 (due September 4th)

Good luck with the semester!

jason.bryer@cuny.edu

angela.lui@cuny.edu

george.hagstorm@cuny.edu

DATA606spring2024.slack.com

@jbryer

@angelalui11

@jbryer@vis.social

spring2024.data606.net