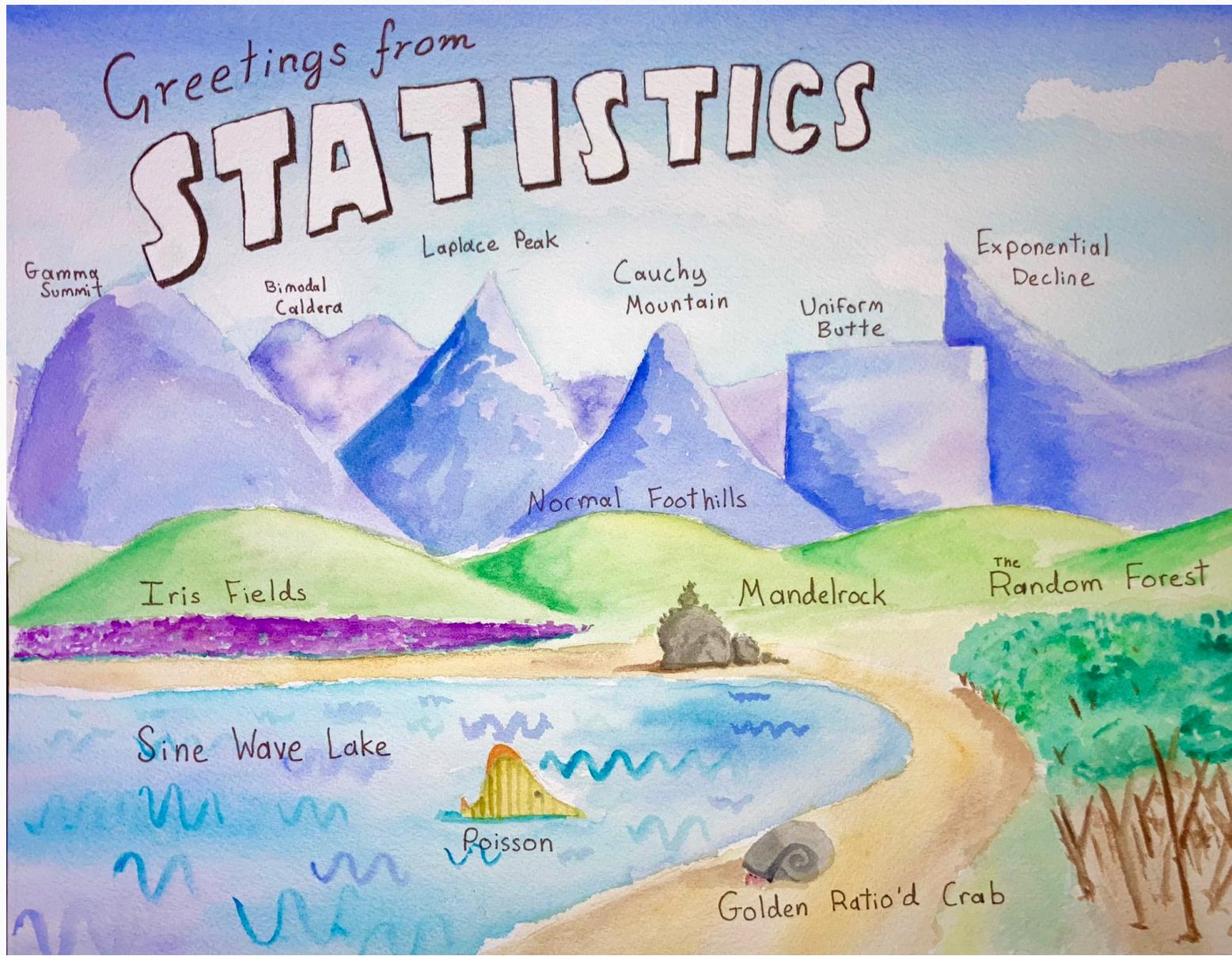


Introduction

DATA 606 - Statistics & Probability for Data Analytics

Jason Bryer, Ph.D.

Spring 2021



@skyetetra

Agenda

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

Introduction

A little about me:

- 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.



Syllabus

Syllabus and course materials are here: <https://spring2021.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://spring2021.data606.net/course-overview/software/>

Meetups

We will have meetups on Wednesday evenings at 8:30pm.

Meetups will be recorded and made available the next day on the [course website](#).

Though attending live is not strictly required, **I 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/gY9SeBCPggHEtZYw6>

Schedule

Start	End	Topic
Friday, January 29, 2021	Sunday, February 07, 2021	Chapter 1 - Intro to Data, R, and Rstudio
Monday, February 08, 2021	Sunday, February 14, 2021	Chapter 2 - Summarizing Data
Monday, February 15, 2021	Sunday, February 21, 2021	Chapter 3 - Probability
Monday, February 22, 2021	Sunday, March 07, 2021	Chapter 4 -Distributions
Monday, March 08, 2021	Sunday, March 14, 2021	Chapter 5 - Foundation for Inference
Monday, March 15, 2021	Sunday, March 21, 2021	Chapter 6 - Inference for Categorical Data
Monday, March 22, 2021	Friday, March 26, 2021	Chapter 7 - Inference for Numerical Data
Monday, April 05, 2021	Sunday, April 18, 2021	Chapter 8 Linear Regression
Monday, April 19, 2021	Sunday, May 02, 2021	Chapter 9 - Multiple & Logistic Regression
Monday, May 03, 2021	Monday, May 17, 2021	Intro to Bayesian Analysis
Wednesday, May 19, 2021	Sunday, May 23, 2021	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 textbook has a chapter on Bayesian analysis that we will use at the end of the semester.

Assignments

- DAACS (6%)
- Participation (10%)
 - One Minute Papers
 - Meetup Presentation - Present one practice problem during our weekly meetups. Signup using the [Google Spreadsheet](#). **Please select odd number questions only!**
- Homework (18%)
- Labs (36%)
 - 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 (20%)
 - 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.
- Final exam (10%)

Communication

- Slack Channel: <https://data606spring2021.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 course or DATA606 package:
<https://github.com/jbryer/DATA606spring2021/issues>
- Email: jason.bryer@sps.cuny.edu
- Phone/Zoom: Please email to schedule a time to meet.
- Office hours will typically be:
 - Fridays from 12:00am to 1:00pm
 - I will use the same Zoom link that we use for the Wednesday night meetups.

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://spring2021.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:

- Complete this Google form: <https://forms.gle/FGsUWy61k8A3ujYH9>
- Create an account at <https://my.daacs.net> and complete the self-regulated learning assessment
- Join the Slack channel

Then:

- Attend the meetup on Wednesday (February 3rd) at 8:30 pm (or watch the recording)
- Start Lab 1 (due February 7th)
- Start Homework 1 (due February 7th)

Good luck with the semester!

 jason.bryer@cuny.edu

 data606spring2021.slack.com

 @jbryer

 @jbryer

 @jasonbryer

 spring2021.data606.net

 bryer.org