431 Syllabus - Fall 2018

Thomas E. Love, Ph.D.

Version 2018-08-06 11:52:36

Contents

Critical Information

Note that this is a **DRAFT syllabus used for testing purposes only**, and will change drastically in the weeks to come. Everything in this document is subject to substantial change before class begins at 1 PM on 2018-08-28 in Room E321-323 in the Robbins Building at the School of Medicine.

This is the Fall 2018 syllabus page for PQHS / CRSP / MPHP 431: Statistical Methods in Biological & Medical Sciences, Section 1, taught by Professor Thomas Love. The course is given on Tuesdays and Thursdays from 1:00 to 2:15 PM, in Robbins Room E321-323.

Course Home Page

The course home page, with links to everything else you'll need, is at https://github.com/THOMASELOVE/431-2018

Getting Help!

To get help for anything related to the course, email **431-help at case dot edu**.

- Dr. Love is available on Tuesdays and Thursdays at CWRU, by appointment. To make an appointment, email him at `thomas.love@case.edu`. He is usually available for 15-30 minutes both before and after class for drop-in conversations. His office is Wood WG-82 L.
  - If you have any special concerns about the course, need special accommodations or any other issues for Dr. Love, please email or stop by before or after class.
- TA office hours will be more extensive, and will begin in early September. A schedule will be announced as soon as possible. Those office hours are held either in the computing lab (WG-56) or the student lounge (WG-67) on the ground floor of the Wood building, depending mostly on the TA's preference that day.

Chapter 1

Course Description

PQHS 431 (cross-listed as CRSP 431 and MPHP 431) is the first half of a two-semester sequence (with PQHS 432) focused on modern data analysis and advanced statistical modeling, with a practical bent and as little theory as possible. We emphasize the key role of thinking hard, and well, about design and analysis in research.

The course is formally titled *Statistical Methods in Biological & Medical Sciences, Part 1.* A more accurate title is **Data Science for Biological, Medical or Health Research**.

We'll learn about managing and visualizing data, building models and making predictions, and other "data science" activities. This highly applied course focuses on modern, more than classical, tools for learning from data. We'll learn a lot of R, and we'll use R Studio and Markdown as tools to help make R work better, and perform our research in replicable ways.

1.1 431 in Three Parts

- Part A is basically August/September and is about Visualizing Data.
- Part B happens in October, and is about Making Comparisons.
- Part C is in November/December. It's about Building Regression Models.

1.2 What will you be doing in the 431-432 sequence?

1. Using modern data science tools to import, tidy/manage, explore (through transforming, visualizing, and modeling) and communicate about data.
2. Thinking hard, and well, about design and analysis in scientific research. We want students to see the value of statistical thinking throughout the process of doing scientific research.
3. Programming in R sufficient to accomplish the tasks above, with enough self-sufficiency afterwards to be able to debug and use new R tools without substantial troubleshooting help from others. What separates "doing data science" from "doing data analysis" is programming.
4. Learning about the importance of replicable research, and developing facility and practice in open source tools (R Markdown, and GitHub, mostly) all the time, as a matter of course.
5. Gaining sufficient background in the practical issues regarding linear and generalized linear models (a big example: missing data) to give you a starting place for meaningful applied work, particularly in terms of making comparisons to address general types of statistical and analytic questions (exploratory, predictive, inferential, and causal, in particular.)

1.3 What do we assume you know before you take the course?

Not much. Useful prior experience includes training/experience in statistics, coding/programming and biology/biomedical science. We expect most people will have some experience in one or two of these areas, but very few have all three.

- Some students have lots of prior training in statistics. But there are many students in the class with no statistical training at all that they use regularly. We assume only that everyone knows what an average is, and has some sense of why statistics might be useful to them in their chosen field.

- Some students have lots of prior coding and programming experience, including experience with R. Some have never written a line of code in their life. We assume only that everyone is willing to learn how to do modern work with data, and that means writing computer code, but that some people will be starting from nothing.
- Some students have lots of prior experience with biological and biomedical science, and know a lot of useful things in those areas which relate directly to our work. Others have zero experience in this area, and will learn a lot from their colleagues. We assume only that everyone is willing to learn, and to put in some effort to do so.

People take this course with a wide range of backgrounds and a common interest in using data effectively in research related to biology, health or medicine. There will be multiple people in the class who are years away from their last statistics class, and the vast majority of students will have no prior experience using R, or any meaningful recollection of using statistical software. The pace can be brisk at times, but all CWRU students who feel up to it are welcome, regardless of their field of study or prior experience.

1.4 What will we learn in 432?

If you have specific questions about 432 not addressed here, just ask Dr. Love.

Chapter 2

Class Schedule

The main course calendar is linked at https://github.com/THOMASELOVE/431-2018. Go there for details on each class throughout the semester.

As mentioned, the course divides neatly into three parts. Classes 2-5 are more about R and R Studio than about statistics. Once everyone's gotten rolling, our approach changes a bit to focus more on statistical concerns, and less on the technology.

2.1 Part A (Classes 2-9) is mostly about R and Visualizing Data.

- Exploratory Data Analysis
  - Descriptive Numerical and Graphical Summaries
  - Distributions, specifically the Normal
  - Histograms and their cousins
  - Scatterplots and related tools from correlation and linear regression
- Exploring Data with the Tidyverse, Getting Up To Speed with R
  - Visualizing Data with `ggplot2`
  - Data Transformation and `dplyr`
  - Using scripts and projects, Building Code

2.2 Part B (Classes 10-17) is about Making Comparisons.

- Estimation and Inference for Means and Proportions
  - Confidence Intervals
  - Design Implications: Matched vs. Independent Samples
  - Hypothesis Testing Strategies
  - Cross-Tabulations
  - The Analysis of Variance and Multiple Comparisons
  - Dealing with Missing Data
  - Randomized Trials vs. Non-Randomized Studies

2.3 Part C (Classes 18-27) is about Building Regression Models.

- Estimation and Inference using Ordinary Least Squares

  - Building Prediction Models, and Validating Them
  - Residual and Influence Analyses
  - Foundations of Model Selection

2.4 Topics Discussed Extensively in 431 and its follow-up, 432

1. Exploratory Data Analysis: "All graphs are comparisons" including data exploration, statistical graphics and more general visualization of information.
2. Placing biological, medical and health research questions into a statistical framework.
3. Study Development - making choices in designing and executing the collection and aggregation of data.
4. Data Handling - including important issues in importing, tidying and transforming data, as well as methods for dealing with missing data, including imputation.
5. Statistical Comparisons: "All of statistics are comparisons" - including methods for discrete and continuous variables: intervals, assumptions, some thoughts on statistical power, and the bootstrap, design of visualizations and models for rates, proportions and contingency tables.
6. The proper use of multi-predictor models for continuous and discrete data, including...
    - Fitting, evaluating, and interpreting linear and generalized linear models.
    - Prediction and validation.
    - Critical role of graphics, including diagnostics and residual analysis.
    - Model choice, including variable selection, shrinkage and model uncertainty.
    - Dealing with categorical predictors and interactions meaningfully.
    - Causal inference using regression: controlling for covariates meaningfully.
7. Using R and R Studio to make all of the things above happen; with particular emphasis on doing replicable research and using Markdown to document the work.

Need more details on course topics? **Dr. Love's 431 book** will be available before the start of class.

Chapter 3

Software

The course makes heavy use of the R statistical programming language. Details on downloading and installing R and the development environment, R Studio, for either PC or Mac, are provided below.

There will be many people in the course for whom R is a new experience. I assume no prior R work in the course. You will know a fair amount of R (and some other things, too) after taking the course, though. We'll also be using the Markdown tool within R Studio. R Markdown will be taught in our class, and can be used to generate reproducible reports that appear as .html files or Word documents, just to give two examples.

3.1 Instructions for Installing R and R Studio

R and R Studio are two different things, but each is free software.

Complete instructions, with a step-by-step walkthrough, are available at https://github.com/THOMASELOVE/431/blob/master/software-installation-431.md

If you need more help, you might look at this terrific resource for Installing R and RStudio from Jenny Bryan and the STAT 545 project. These are the people responsible for the great Happy Git with R project, which is worth your time, too, if you intend to use Git and GitHub. (Everyone will in 432.)

In brief, the steps you need to take for 431 are:

1. Download and install the latest version of R (version 3.4.1 or later) at http://cran.case.edu/ or https://cran.r-project.org/.
2. Download and install the preview version of R Studio (version 1.1.345 or later) at https://www.rstudio.com/products/rstudio/download/preview/.
3. Install some R packages - an R "package" is a collection of functions, data, and documentation that extends the capabilities of R, and is the critical way to get R doing interesting work. To install the packages for our course, open R Studio and run these commands.

```
pkgs <- c("aplpack", "arm", "babynames", "boot", "car", "devtools", "Epi",
          "faraway", "forcats", "foreign", "gapminder", "GGally", "ggjoy",
          "gridExtra", "Hmisc", "knitr", "lme4", "markdown", "MASS",
          "mice", "mosaic", "multcomp", "NHANES", "pander", "psych",
          "pwr", "qcc", "rmarkdown", "rms", "sandwich", "survival",
          "tableone", "tidyverse", "vcd", "viridis")

install.packages(pkgs)
```

3.2 Getting Started with the Software, once you've installed

1. Dr. Love's document Getting Started with R might be a good first step. This is basically a demonstration of how to use these tools to actually analyze data.
2. Dr. Love also prepared a downloadable template for your first few R Markdown attempts. Get it by downloading the data and code for the course at https://github.com/THOMASELOVE/431data. Click on the green Clone or download button, and then select Download ZIP to obtain a Zip file of all posted materials.
3. We can also recommend Chester Ismay's Getting Used to R, RStudio and R Markdown as an introduction to the basics.
4. Dr. Love will demonstrate the use of R, R Studio and R Markdown in class, starting with Class 2.
5. Dr. Love's Course Notes are a source of many examples.

3.3 Why do we teach R, instead of SPSS or SAS or whatever, in 431-432?

Because it is by far the better choice for what we're trying to do, which is to help you become effective data scientists. And effective scientists, period.

Chapter 4

Texts

4.1 Dr. Love's Notes

The main text is a set of Notes for the course, maintained by Dr. Love at https://thomaselove.github.io/431notes/.

Although the Notes share some of the features of a textbook, they are neither comprehensive nor completely original. The main purpose is to give 431 students in Section 1 (and Section 2) a set of common materials on which to draw during the course, providing a series of examples using R to work through issues that are likely to come up during the semester. The material will be updated regularly as the semester progresses.

Slides from each session of the class are posted as .pdf files at https://github.com/THOMASELOVE/431slides

4.2 Books To Purchase

In addition, we'll read two books that you'll need to purchase (the combined price is about $25.):

1. Nate Silver's The Signal and The Noise ISBN-13: 978-1594204111 Amazon Link, and
2. Jeff Leek's The Elements of Data Analytic Style, available at https://leanpub.com/datastyle.

With regard to The Signal and the Noise, you can watch Nate discuss the book's ideas in many places, for instance, at this YouTube link, or this one on the Art and Science of Prediction, or this one at Google. We'll also spend considerable time (even before we read the book) looking at some articles from the FiveThirtyEight website, where Nate is editor-in-chief.

4.3 Free Resources You'll Definitely Need To Access

4.3.1 Textbooks

1. OpenIntro Statistics (OpenStats) by David Diez, Christopher Barr and Mine Cetinkaya-Rundel. This is an excellent resource, with lots of useful information set at a reasonably elementary level.
    - In Part A of the course, you'll want to look at Chapters 1 and 3, in particular.
    - Part B: Chapters 4, 5, 6
    - Part C: Chapters 7, 8
2. R for Data Science (R4DS) by Garrett Grolemund and Hadley Wickham - this is a great resource, but may feel a little advanced for those of you brand new to coding, who may want to supplement it.

- In Part A, we'll discuss ideas from the Introduction and Explore sections, mostly.
- Parts B and C will address some issues discussed in the Wrangle, Model and Communicate sections.

3. Practical Regression and ANOVA using R, by Julian J. Faraway, (Faraway) which is one of the "More Free Books" to download at https://www.openintro.org/stat/extras.php. Also uses R, but much more focused on statistical issues. A more formal presentation is in Linear Models with R, Second Edition by Julian J. Faraway (Chapman and Hall / CRC Texts in Statistical Science) ISBN-13: 978-1439887332. But the free text is sufficient for 431 and, probably 432.
    - Faraway's material is mostly a good resource for Part C, although Chapter 16 will help with ANOVA in Part B.

4.3.2 Articles

1. Several of the guides prepared by Jeff Leek and his group, including:
    - Finally, a Formula for Decoding Health News, from fivethirtyeight.com
    - How to share data with a statistician,
    - Reading academic (scientific) papers,
    - Writing your first academic paper
    - Write papers like a modern scientist
2. Part of the Ten Simple Rules series at PLOS Computational Biology, specifically
    - Ten Simple Rules for Effective Statistical Practice by Kass RE et al. 2016
    - Ten Simple Rules for Graduate Students by Gu J Bourne PE 2007
    - Ten Simple Rules for Better Figures by Rougier NP Droettboom M Bourne PE 2014
    - Ten Simple Rules for Creating a Good Data Management Plan by Michener WK 2015
    - Ten Simple Rules for Reproducible Computational Research by Sandve GK et al. 2013
3. The American Statistical Association's Statement on p-Values: Context, Process and Purpose
    - We'll also look at some of the Supplemental Material.
4. The preprint from Benjamin D Berger J Johannesson M et al. called "Redefine statistical significance", which proposes to change the default p-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005. The manuscript will eventually appear in the journal *Nature Human Behavior*.
    - Kelly Servick's related article "It will be much harder to call new findings 'significant' if this team gets its way" from Science, 2017-07-25, comments on the preprint.

4.4 Supplemental (and Free) Texts That May Be Worth Your Time

1. Ismay C Getting Used to R, RStudio and R Markdown - designed to provide new users to R, RStudio, and R Markdown with the introductory steps needed to begin their own reproducible research.
    - We recommend you use this material to help understand some of the basics of these three software tools. Use other sources to supplement statistical content.
2. Ismay C Kim AY ModernDive: An Introduction to Statistical and Data Sciences via R - intended to be a gentle introduction to the practice of analyzing data and answering questions using data the way data scientists, statisticians, data journalists, and other researchers would. Some nice material for all three Parts of our course.
    - In Part A, you'll be looking at the Data Exploration via the Tidyverse materials in this text.
    - In Part B, we'll definitely be looking at the Inference materials.
    - Part C expands on what's in the Data Modeling using Regression section.
3. Horton NJ Pruim R Kaplan DT *A Student's Guide to R* from Project MOSAIC. Most recent updates (pdf) at this link - you may need to *scroll down*. Free, downloadable PDF - an excellent guide to Getting Started with R Studio, and then working through some straightforward examples of how to deal with data in R. Makes heavy use of the `mosaic` package.

- Part A of our course discusses ideas from Chapters 3, 13, 15 and some of Chapter 5.
- Part B discusses ideas shown in Chapters 4, 6, 7 and 12.
- Part C discusses Chapter 5 and 8, and some of Chapter 10.

4. Peng RD Exploratory Data Analysis with R - especially useful material on using R for graphics and general EDA strategies. Covers some basic principles of constructing informative graphs.
    - In Part A, Chapters 3-6 may be helpful. The Case Study in Chapter 16 is interesting, and has a related video.
5. Peng RD R Programming for Data Science - designed to help you get started with the basics of the language, learn how to manipulate datasets, how to write functions, and how to debug and optimize code, which will be more of an issue for us as the semester progresses.
    - Covers some of the same ground as the other Peng book, but at a level geared more for programming.
6. Harrell FE Biostatistics for Biomedical Research - this is more a set of course notes than a full-fledged book, and uses R but not R Studio, mostly. However, it's full of great, in-depth information on basic statistical methods, and likely to be very useful for Parts B and C of the course.
    - Chapters 1-3 include introductions to relevant R, algebra and biostatistics.
    - For Part A, the value is in Chapter 4 and some of Chapters 14 and 21.
    - Part B - see Chapters 5-7.
    - Part C - consider Chapters 8-12.
7. R Studio has great Cheat Sheets for Data Import, Data Transformation, Data Visualization, R Markdown and other topics at https://www.rstudio.com/resources/cheatsheets/

Chapter 5

Videos, Podcasts and Other Resources

5.1 Videos to Learn R

Lots of people like to watch a video to learn things. Here are some of the many R examples we've found useful.

1. Getting Started with R Markdown, from R Studio.
2. The Datacamp Course called R for the Intimidated is useful, too, especially for those feeling that way.
3. Getting Your Data into R, from R Studio.
4. Data Wrangling with R and RStudio, also from R Studio.
5. Also good is the six part series on R Studio Essentials, from R Studio.
    - Parts 1 and 2 of each section (Programming, and Managing Change) may be of more interest.

5.2 Podcasts

1. Not So Standard Deviations: The Data Science Podcast with Hilary Parker and Roger Peng talking about the latest in data science and data analysis in academia and industry.
2. The Effort Report with Elizabeth Matsui and Roger Peng talking about life in the academic trenches, telling it "like it is". Every graduate student in this course looking at a career in academia would benefit from listening.

5.3 A Few Web Sites and Blogs

1. FiveThirtyEight
2. Andrew Gelman's Blog: Statistical Modeling, Causal Inference and Social Science
3. Kaiser Fung's Junkcharts Blog
4. Nathan Yau's Flowing Data Blog

There are several curated lists of data science blogs online, for example, see https://github.com/rushter/data-science-blogs

5.4 Other Resources

1. The `swirl` package in R can be a great help for people learning R programming and data science. Find out more about it at http://swirlstats.com/students.html
2. UCLA's Institute for Digital Research and Education has some great Data Analysis Examples using R (and other software.)

Chapter 6

Dr. Love

Thomas E. Love, Ph.D.

- Professor of Medicine, Population and Quantitative Health Sciences, CWRU
- Director of Biostatistics and Evaluation, Center for Health Care Research & Policy, MetroHealth Medical Center
- Chief Data Scientist, Better Health Partnership
- Track Lead for Health Care Analytics, MS in Biostatistics, Department of Population and Quantitative Health Sciences, CWRU
- Fellow, American Statistical Association

Email

- Email to get help with the course: **431-help at case dot edu** (seen by Professor Love and the TAs)
- Thomas dot Love at case dot edu (for matters related to grades or individual concerns)
- Dr. Love is hard to reach by phone. Email is always the best way to reach him.

Offices

- Wood WG-82L on the ground floor of the Wood building (Tuesdays and Thursdays)
- Rammelkamp R-229A at MetroHealth Medical Center (Wednesdays and Fridays)

Dr. Love is generally available before and after class, otherwise by appointment on Tuesdays and Thursdays (send email to schedule).

Web

- Dr. Love's GitHub pages website.
  - His GitHub name is THOMASELOVE.
- His Twitter handle is [@ThomasELove](https://twitter.com/ThomasELove).

A More Complete Biography

Hi. I am Thomas E. Love, Ph.D. and I have at least three different jobs.

- I am a Professor in the Departments of Medicine and Population & Quantitative Health Sciences at Case Western Reserve University. I teach three courses per year there (PQHS 431, 432 and 500) and also lead the Health Care Analytics track of the MS program in Biostatistics.
- I direct Biostatistics and Evaluation at the Center for Health Care Research & Policy, which is a joint venture of CWRU and MetroHealth Medical Center.
- For ten years, I was the (founding) Data Director for Better Health Partnership, an alliance of people who provide, pay for and receive care in Northeast Ohio. I now serve as Chief Data Scientist there.
- I am a Fellow of the American Statistical Association, and have won numerous awards for my teaching and my research.
- I have been teaching at CWRU since 1994, and have taught every manner of CWRU student over the years, especially students in biostatistics, medicine, and management.

In research, I use statistical methods to look at questions in health policy and in particular the provision of health services. I mostly work with observational data, rather than data that emerge from randomized clinical trials, and I have a special interest in working with data from electronic health records.

- You may be interested in a recent study in Health Affairs showing the impact of a Medicaid-like expansion plan on care and outcomes of poor patients in Cleveland.
- Or you might be interested in our New England Journal of Medicine study of the effect of electronic health records on the care and outcomes of people with diabetes.
- In 2011, James O'Malley and I chaired the Ninth International Conference on Health Policy Statistics, here in Cleveland. Here's a recap. We may chair it again in 2021.
- I've also worked on many projects involving the use of propensity scores to make causal inferences from observational studies, particularly in heart failure.

If you want to see a pretty complete list of my publications, knock yourself out.

I hold degrees from Columbia University in the City of New York and from the University of Pennsylvania. My dissertation advisor was Paul Rosenbaum. I am married to a brilliant woman and we are raising two terrific sons, one of whom just started college. I live in Shaker Heights.

Chapter 7

Teaching Assistants

The teaching assistants for 431 this year will be determined this summer. They are the people answering 431-help at case dot edu, and they are the people holding the bulk of our office hours. Each of them has been in your shoes - they've taken the course in the past, and they enjoyed it enough to come back for more. Many have volunteered their precious time and energy to help make the course happen, and we couldn't be more delighted to welcome you to the course.

To contact the TAs, email `431-help at case dot edu`.

7.1 Office Hours for TAs

Teaching Assistant Office Hours are held in WG-56 (Computing Lab) or WG-67 (Student Lounge) on the ground floor of the Wood building, so be sure to look in both places if you need help.

TA office hours will be specified in early September.

Chapter 8

Expectations, Assessment and Grades

All students are expected to attend all sessions, participate vigorously in the class discussions and in team work, complete all individual work in a timely fashion (**late work will be accepted only if approved by Professor Love, and he will only grant that approval in catastrophic circumstances**), demonstrate improvement of skills over the term, and perform well on the Quizzes and in the final portfolio presentation. Such a performance is the minimum standard required to receive a grade of B

To receive an A, students are expected to complete all the requirements described above, demonstrate excellent work in both the final portfolio presentation, and outstanding work in at least one of the following: [a] in class participation, [b] assignments, [c] quizzes.

Grading standards apply in the same way for all students, regardless of whether they are enrolled in PQHS 431, CRSP 431 or MPHP 431. The courses are identical.

The course grade is based on four key areas of demonstrated accomplishment. The planned breakdown is as follows, but Dr. Love may make adjustments as the semester progresses.

Weight	Task
15%	In-Class and Outside-of-Class Participation
25%	Completion and Quality of Assignments 1-6
30%	Peformance on Quizzes 1-3
30%	Final Portfolio Presentation and Related Materials

Any questions regarding how you are doing in the course can be directed to Dr. Love.

8.1 Participation

Students often ask how they can improve this part of their grade. I cannot emphasize enough how much we want to hear from you about things that are relevant to this course.

1. If you're not shy, ask questions in class. The TAs help me assess participation, so they are paying attention, too. Come to the TA office hours if you need help. Make an appointment to talk to us if you have something to discuss that doesn't work well in email.
2. Email `431-help at case dot edu`. with your questions and comments. That'll lead to faster answers, typically, and help us recognize you as someone trying to improve their understanding.
   - Find **typos** in the materials (code, slides, the Notes, this syllabus)? Send them to us at `431-help at case dot edu`.

- See a cool visualization online? A nice use of statistical methods or design in a paper? Share them with us, at `431-help at case dot edu`.

3. Mindlessness and poor planning are unimpressive. Be smart and **plan ahead**. Do the work on your end first. In the last 18 hours before something is due, we generally do not answer questions, deliberately.

Chapter 9

Quizzes

There are three quizzes scheduled. Each Quiz will be taken online, exclusively.

- **If you need to make alternate arrangements for a Quiz, please contact Professor Love via email as soon as possible**, and in any case, at least a week before a Quiz is released.
- A test survey will be completed in September to ensure that you can use the quiz software to submit your results.
- You will have a minimum of 91 hours to work on each quiz. Most students complete each quiz in 3-5 hours.
- All quizzes require an Internet connection, and a CWRU login.
- Late submissions of Quizzes will not be accepted.
- Quizzes typically involve somewhere between 30 and 50 short-answer questions.

9.1 A few general comments on the quizzes

1. The questions are not arranged in any particular order, and you should answer all questions.
2. All questions involve relatively short responses, sometimes after working through a detailed analysis.
3. You will have the opportunity to edit your responses after completing the Quiz, but this must be completed by the deadline.
4. You are welcome (even encouraged) to consult the materials provided on the course website, but you are **not** allowed to discuss the questions on the Quizzes with anyone other than Professor Love or the teaching assistants.
5. We do not guarantee to answer questions we receive via email less than 3 hours prior to the Quiz submission deadline.
6. After each Quiz is complete, an answer sketch will be made available. Grades for Quizzes 1 and 2 should be available within 36 hours of the due date, and for Quiz 3, they will be available by Friday 2017-12-15.

Chapter 10

Assignments

There are six main homework assignments scheduled. Most require straightforward demonstrations of mastery for core principles and fundamental skills. Some require deeper dives into more technically sophisticated material. Some also require reflection, particularly based on materials we'll be reading throughout the semester, especially from Nate Silver's book.

1. Assignment 1 is due Friday 2017-09-15 at noon.
2. Assignment 2 is due Friday 2017-09-22 at noon.
3. Assignment 3 is due Friday 2017-09-29 at noon.
4. Assignment 4 is due Friday 2017-10-27 at noon.
5. Assignment 5 is due Thursday 2017-11-09 at noon.
6. Assignment 6 is due Monday 2017-12-04 at noon.

10.1 Where will I find the Assignments?

The actual homework assignments are found at https://github.com/THOMASELOVE/431homework

10.2 Where do I turn in the Assignments?

You will turn in your Assignments using the Canvas system at https://canvas.case.edu.

The course's primary listing is PQHS 431, but students in CRSP 431 and MPHP 431 should find the same information. The link to post your responses for Assignment 1 will be there by the first day of class. Subsequent assignments will appear after the deadline for the preceding assignment has passed.

10.3 General Comments on Assignments

1. Each assignment will require you to analyze some data, and prepare a report using R Markdown. You will submit both your Markdown file, and an HTML, PDF or Word document built from that Markdown file.
2. Most assignments will require you to write an essay. Essays must be composed as part of your Markdown file, and thus included in your HTML/PDF/Word document. Do not edit the result of your R Markdown conversion into Word.
3. When writing in English, use complete sentences, rather than bullet points.

4. Clearly mark each Question in each Assignment. There is no need to repeat the question before answering it, although you are welcome to do so.
5. Read and heed the advice of Jeff Leek in *The Elements of Data Analytic Style*. Chapters 5, 9, 10 and 13 of that book are especially relevant to our early assignments.
6. You are welcome to discuss each Assignment with anyone, including Dr. Love, the teaching assistants, or your colleagues, but your answer must be prepared by you alone. We especially encourage you to take advantage of TA office hours and email 431-help at case dot edu. 7, In general, we do not provide answers to questions that we receive in the last 18 hours before an assignment is due. So don't leave anything until the last day. Allow time for computer problems.
7. Late work is inappropriate for graduate school. Failure to turn in an assignment within one hour of the deadline (all deadlines are noon) will result in a very poor grade on the assignment if it is (eventually) turned in, and a zero (from which it is difficult to recover) if it is not turned in. Submission of timely, but partial work is far better than no submission at all.
8. If an assignment is scheduled so that you will not be able to complete it in a timely fashion, it is your responsibility to email Dr. Love about the situation so he can evaluate it. Such requests should be sent as soon as possible, and at least 48 hours prior to the deadline for the assignment, except, of course, in the case of truly horrific circumstances.
9. Grades on Assignments are usually available one week after the submission deadline.

Chapter 11

Final Portfolio Project

All materials and information related to the project are maintained at https://github.com/THOMASELOVE/431project and regular updates will appear there throughout the semester.

Chapter 12

A Few Writing/Presenting Tips

1. Statistics is a "getting the details right" business - we care deeply about details, and this applies to writing code or complete English sentences.
2. Nothing impresses us as much as a clear and concise argument, presented using well-written English sentences, effective and well-labeled figures and tables.
3. Don't parrot back material that Dr. Love wrote or said. State ideas in your own words. Stating them in other words is, technically, plagiarism.
4. Edit your more adventurous output; don't present everything you know how to do in R, and don't forget that someone is trying to read both your code and your results.
5. Make your work easy to evaluate. In responding to an assignment, be sure to answer the question that was asked, restating it as necessary.
6. Clearly label everything: graphs, tables, your answer to a specific question. Everything. Again, make your work easy to evaluate.
7. Simplify. Emphasize ideas in plain language. Avoid jargon. Use English well.
8. Data are plural. Use "the data **are** ... " rather than "the data *is* ... "
9. A paragraph must contain more than one sentence.

10. Don't switch tenses. If you want to write in the present tense, stick to it throughout.
11. Don't write or say random sample unless you used a random number generator. If you used haphazard sampling or convenience sampling, call it what it is, and indicate whether any problems could have cropped up as a result.
12. Similarly, don't defend a method of data collection because it is random. Most of the time we want to represent some population, and a random sample is just one way to ensure that certain types of biases have a low probability of creeping in.
13. If you want to write that you used $\alpha = 0.05$ as your significance level, then state that your results were obtained using a 95% confidence level, not a 95% confidence interval, unless you are actually interpreting a confidence interval.

14. If you're looking at a p-value, then you should state either:
    - [1] We're using a 95% confidence level.
    - [2] We're using a 5% significance level. or
    - [3] We're using $\alpha = 0.05$.
    - Don't use more than one of these expressions. 15 Refer to all p-values that are less than 0.001 or perhaps less than 0.0001 as $p < 0.001$, rather than, for instance, $p = 0.00000001$ or, worse yet, $p = 0$. In a similar vein, write all p-values that exceed 0.99 as $p > 0.99$ instead of, for instance, $p = 1$.

15. To the extent possible, don't use `computer-ese` to label variables, plots or tables. R and Markdown allow you to change the labels on graphs and tables to meaningful things – do so. Use meaningful abbreviations, as necessary, explaining what they mean on the first usage.
16. When in doubt, err on the side of clarity. Clear thinking, clear writing.

Chapter 13

General Course Policies

1. Any concerns or questions regarding these general policies, the teaching assistants or the course itself should be directed to Dr. Love, if at all possible.
2. All student work is subject to the University's policies and procedures.
3. **Registration is required**. I do not permit anyone to audit the course, without exception.
4. **Grading**. You are not in competition with each other for grades. I have no set percentage of students who will receive any particular grade.
5. **Attendance** is expected, and your absence will be noted. If you need to miss class, inform Dr. Love via email before the class you will miss, or as soon as possible thereafter. I will assume you have a good reason - details are **not** necessary. You are responsible for all missed work, regardless of the reason for your absence. All work is turned in electronically, except for the final project presentation.
6. **Late work is unacceptable** under anything but the most harrowing of circumstances. Dr. Love (via email) is the person to discuss this with, at least 48 hours prior to the deadline, if you feel your circumstances are sufficiently dire to warrant an exception. It is far better to turn in timely, but only partially complete work than nothing at all.
7. **Feedback on assignments - deadline.** On every assignment, Quiz, project-related task, whatever, we will be delighted to respond to email questions **up to 18 hours before** the assignment is due. After that time, you are on your own. The reason for this is that Dr. Love and the teaching assistants will regularly post responses to frequently asked questions about assignments, and we need sufficient time to accomplish this task.
8. **On Getting Help Quickly and Effectively**: In general, we don't have a way to diagnose your problem with R, R Studio or Markdown if you don't show us what you're typing that causes an error, or a lack of results. If you wrote a Markdown file, send it, along with a specific question (or series of them) about specific error messages or strange results you are getting. We need to replicate the problem in order to know how to fix your problem, and it also helps if we know what error message you're seeing, or what strange result you are getting.
9. **Using a Laptop** Using a laptop to follow along, take notes, or try things out during class, can be very helpful. Feel free to do so.
10. **Computer** You will need access to a computer (PC or Mac - a ChromeBook won't do) outside of class to do every assignment. You need to be able to install software on this computer, and update it frequently.
11. **Distractions**. Silence your phone during class. The temptation to look at your phone or Facebook or email during class is nearly irresistible. Resist anyway, if only to avoid distracting your instructor and your fellow students. **Dr. Love has absolutely no shame about embarassing people on this issue. If it's critical, just step out of the room**.
12. **Research Usage**. Any and all results of in-class and out-of-class assignments and activities are data sources for research and may be used in published research. All such use will always be anonymous.
13. **Audio-Recording.** Anything you say during a class session *may* be audio-recorded.
14. **Typos.** Dr. Love makes occasional typographic and grammatical errors, which irritate him enormously.

Please email him if you find any in this syllabus or any other course materials. If you are the first to let us know, and we make the change, you will receive a small amount of bonus credit in your class participation grade.

13.1 Grade Appeal Policy - Wait until December!

For each assignment and Quiz, after it has been submitted, we will publish a detailed answer sketch and a grading rubric. You will also learn your scores on each individual item.

Clarification of concerns related to potential typographical or other errors in these answer sketches is welcome at any time, but haggling over points on assignments and quizzes can be a real time sink in a large class.

- To that end, students are **requested not to dispute** any grading until December.
- Students are permitted to ask questions about grading during the term, and the teaching assistants and I are happy to discuss why points were taken off, but we make it clear that no grades will be changed until the end of the term. The one exception is if there is a mistake in adding up points, or some similar clerical error. Those are corrected immediately, of course.
- In early December, we will provide a form listing all completed Assignments and Quizzes. Any student who wishes to dispute points can specify the number of points in question next to each relevant assignment or Quiz, and the details of the issues that concern them.
- If you wish to dispute a grade, just fill out the form in December. All forms must be submitted by the end of the final project presentation on December 14.

In mid-December, **after** Dr. Love has worked out what letter grade to give each student, he will go through the requests and determine for each whether the student's letter grade would change if all of the points in dispute were granted. If the answer is no, then we don't even look at the disputed grade(s). If the answer is yes, then we look very carefully to see if enough extra points are merited to change their grade. (It will not help your case if you submit any frivolous requests.)

The main advantage of this system is that it saves all of us (you, the TAs, and Dr. Love) the hassle of haggling over points that are never going to mean anything anyway. It also provides "equal access" to students who are too timid to approach us in person with their concerns. Finally, if there is an issue with grading a particular problem or assignment that needs to be reconsidered, Dr. Love will have access to all papers and can make a universal decision[1]

[1] I got this idea from Jessica Utts at http://www.amstat.org/publications/jse/v22n2/rossmanint.pdf.

Chapter 14

Necessary?

But within those Parts, we jump around. A lot.

An outline of the course schedule follows. The more detailed calendar is at

This is only a **tentative schedule**, but we'll wind up in the same place.

Class	Date	Topics	Readings
1	2018-08-28	Introduction, Getting Started	Syllabus
2	2018-08-30	**Part A. Visualizing Data**, R and R Studio	
3	2018-09-04	Exploring Data and the Tidyverse	
4	2018-09-06	Scripts and Projects in R/R Studio	Leek 5, 9, 10, 13
5	2018-09-11	Data Transformation and dplyr	Silver, Intro and Chapter 1
6	2018-09-13	Exploratory Data Analysis	
-	-	Assignment 1 due at noon.	
7	2018-09-18	Descriptive Statistics, The Normal Distribution	
8	2018-09-20	Using `ggplot2` more effectively	
-	-	Assignment 2 due at noon.	
9	2018-09-25	Linear regression and correlation	Silver, 2-3
10	2018-09-27	Studying associations	
-	-	Assignment 3 due at noon.	
11	2018-10-02	**Part B: Making Comparisons**	Leek 1-4 and 12
12	2018-10-04	Confidence intervals for Means	Quiz 1 provided to you.
-	-	Quiz 1 due at noon.	
13	2018-10-09	Matched Pairs, Independent Samples	Leek 6, Silver 4-5
14	2018-10-11	Hypothesis Testing Strategies	
-	-	Project Task A due at noon.	
15	2018-10-16	The Analysis of Variance and Related Tools	Silver 7-8
16	2018-10-18	Multiple Comparisons	
-	-	Project Task B due at noon.	
–	*2018-10-23*	*CWRU Fall Break (no class)*	–
17	2018-10-25	Comparing Two Means - Set Up Survey	
-	-	Assignment 4 due at noon.	
18	2018-10-30	ANOVA - Comparing Multiple Means	
19	2018-11-01	**Part C: Building Effective Models**	Silver through 11
20	2018-11-06	Building Inferences for Rates	
-	-	Project Task C due at noon.	
-	-	Assignment 5 due at noon.	
21	2018-11-08	Conclude Part B of course	Quiz 2 provided to you.

Class	Date	Topics	Readings
-	-	Quiz 2 due at 8 AM.	
22	2018-11-13	Begin Part C of course	Leek 7-8
23	2018-11-15	Estimation and Inference	
-	-	Project Task D due at noon.	
–	*2018-11-20*	*Class is cancelled.*	–
–	*2018-11-22*	*CWRU Thanksgiving Break (no class)*	–
24	2018-11-27	Residual and Influence Analysis	Silver 12
25	2018-11-29	Prediction and Validation	
-	-	Assignment 6 due at noon.	
26	2018-12-04	More Modeling Content	Silver finish book
27	2018-12-06	Looking Back, and Forward	Quiz 3 provided to you.
-	-	Quiz 3 due at noon.	–
-	-	Project Task E due at noon.	

Final Portfolio Presentations will be held on

- Monday 2018-12-10
- Tuesday 2018-12-11, and
- Thursday 2018-12-13

Project presentations will be scheduled as part of Project Task A, and you will know the date and time of your portfolio presentation by 2017-10-18.