MATH/STAT 209: Introduction to Statistical Modelling, Spring 2018

Tuesday, Thursday 10:30-11:45	JPSN G23
Tuesday, Thursday 12:00-13:15	JPSN G28
Tuesday, Thursday 15:00-16:15	JPSN G23

Instructor:	**Taylor Arnold**
E-mail:	tarnold2@richmond.edu
Office:	Jepson Hall, Rm 218
Office hours:	Tuesday, Thursday 16:15-17:00

Description:

This course broadly covers the entire process of collecting, cleaning, visualizing, modeling, and presenting datasets. It has a MATH designation but is not a *mathematics* course. The focus is on applied statistics and data analysis rather than a detailed study of symbolic mathematics. By the end of the semester you will feel confident collecting, analyzing, and writing about datasets from a variety of fields. You will be able to use these skills to address data-driven problems in a wide range of application domains.

Computing:

To facilitate your ability to actually *do* statistics, nearly every class meeting will involve some form of computing. No prior programming experience is assumed or required.

We will use the **R** programming environment throughout the semester. It is freely available for all major operating systems and is pre-installed on many campus computers. You can download it and all supporting files for your own machine via these links:

```
https://cran.r-project.org/
https://www.rstudio.com/
```

The lab computers in Jepson are available and contain all of the required software. I strongly recommend, however, downloading these on your own machine so that you will be able to work on assignments without needing to work only in the computer lab.

Course Website:

All of the materials and assignments for the course will be posted on the class website:

```
https://statsmaths.github.io/stat209
```

The website contains notes, exam dates, assignment details, and supplemental materials. At the end of the semester, this version of the course will be archived and available for your reference.

GitHub:

Your work for this semester will be submitted through GitHub, the same platform that hosts our website, using the GitHub classroom program. You will need to set up a free account, which we will cover during the first week of class.

Learning Objectives (summary):

The following are twelve learning objectives that will be covered throughout the semester. Each topic will be the focus of one week of the course, with weeks 7 and 14 devoted to reviewing prior material. You will be able to demonstrate mastery of each objective via in-class examinations.

1. reproducible computing
2. variable types and numeric summaries
3. the grammar of graphics
4. advanced graphics
5. filtering and summarizing data
6. statistical inference
7. communicating statistical results
8. normalized data
9. joining relational data
10. working with strings
11. text mining
12. ethical guidelines for statistical practice

Additionally, the following four learning objectives cover forms of presenting results from statistical analyses. These topics will be folded into the semester at relevant points.

13. data dictionary with collected data
14. inferential data analysis with collected data
15. statistical argument with retrospective data
16. exploratory data analysis with gathered text data

You will demonstrate these objectives via written data reports.

Labs:

Most class meetings will have an interactive lab associated with it. These consist of a set of questions that must be answered with either small snippets of code or short descriptive answers. You are responsible for finishing the lab prior to the next class meeting and uploading your answers to GitHub.

First Assessment (objectives 1-12):

On nearly every Tuesday, there will be a short quiz covering the material from the prior week. Note that this includes both a conceptual understanding of the topics covered as well as the ability to apply these concepts to data with code. Passing this assessment yields a passing score for the respective learning objective.

Second Assessment (objectives 1-12):

There will also be two chances for reassessment throughout the semester. These are currently scheduled on the following two class meetings (in the event of unforeseen circumstances, these dates are subject to change):

- 2018-03-01 (Week 07)
- 2018-04-19 (Week 13)

You will be able to answer a group of questions related to learning objectives that were not successfully passed in the first assessment. At a maximum, you will be able to reassess only 4 of the 6 learning objectives in each half of the semester.

Data Reports (objectives 13-16):

In order to demonstrate mastery of learning objectives 13-16, you will also complete four data-oriented reports. These reports are short written documents that mix code, graphics, and prose to provide a comprehensive analysis of a data set. The tentative due dates and topics for these reports are:

- 2018-02-15 (Week 05): data dictionary with collected data
- 2018-03-22 (Week 09): inferential data analysis with collected data
- 2018-03-29 (Week 11): statistical argument with retrospective data
- 2018-04-26 (Week 14): exploratory data analysis with gathered text data

Reports that make a genuine effort to address the task at hand but fail to meet all of the guidelines will be given detailed feedback. There will be a short re-submission window in which modified reports can be re-handed in for credit.

Final Grades:

I believe that grades are meant to indicate how well students were able to demonstrate their achievement of well-defined learning goals. As described in the prior sections, all of the work for this course will be graded as either a passing grade (P) or a failing grade (I = insufficient). Your final grade is generally determined by the following:

- **A** Passing grades on 16/16 learning objectives
- **A-** Passing grades on 14/16 learning objectives
- **B+** Passing grades on 12/16 learning objectives
- **B** Passing grades on 10/16 learning objectives
- **B-** Passing grades on 9/16 learning objectives
- **C** Passing grades on 8/16 learning objectives

To pass the course, you must also (1) miss no more than four class meetings and (2) miss no more than four lab assignments. Failing to fulfill any of these requirements or achieving fewer than 8 of the learning objectives may result in a failing grade for the course.

Class Policies:

The following class policies address some of the most common questions and concerns that students have. If anything is unclear, please feel free to contact me for clarification at any point in the semester.

- **Academic honesty:** Cheating and plagiarism are grave scholarly offenses and potential grounds for expulsion; they are also a major barrier to your intellectual development. You are expected to familiarize yourself with the entirety of the University of Richmond's Honor Code. If you are confused or unsure about appropriate citation protocol or any other aspect of the Honor code, please consult me before turning in an assignment.
- **Special approval:** If you have special approval forms for extra time on exams or any other circumstances I should know about, please speak with me as early as possible so that we can best accommodate your needs.
- **Late work:** You are expected to submit all work on-time. Late labs and reports will not be considered; always hand in something even if it is not perfect.
- **Attendance:** You are expected to both attend and participate in most class meetings. If you must be absent due to illness or other pressing need, please let me know by email as soon as possible. A habit of arriving late is considered equivalent to an absence.
- **Make-up work:** There are generally no make-up examinations. In instances where students have a valid excuse for missing a quiz or examination, please arrange to meet with me as soon as possible. I will assess mastery of learning objectives by an oral examination.
- **Class conduct:** During class I expect you to refrain from checking email, being on phones, or working on assignments for other classes.
- **Computers:** During programming assignments started in class, I expect you to use the computers in the lab. This is helpful for several reasons: it reduces distractions from iMessages and other materials on your laptop; all of the lab computers are configured using the same software and language set-up, reducing errors specific to your machine; and, other students and myself can share the same screen without worrying about modifying something on your personal machine.
- **Office hours**: If you would like to meet during my office hours, please just come by. No need to schedule an appointment. If you find me in my office at other times of the week, I am usually glad to meet then as well. Finally, I am also happy to make appointments outside of my normal office hours. These appointments are meant for discussing longer issues that are not appropriate for regular office hours (i.e., asking for recommendation letters or discussing an extended absence) or for students who cannot make my normal office hours. Please note that appointments should be booked at least 24 hours ahead of time.
- **Email:** I will also answer questions by email (it can, in fact, be much faster than scheduling an appointment for small issues). During the week, I aim to respond within 24 hours, with emails sent over the weekend responded to by Monday morning. If your question involves code, please attach your current lab or report as that will expedite my answering your question(s).

Learning Objectives (detail):

Below is a more detailed description of each week's material. Note that the exact topics, pace, and coverage may change over the course of the semester.

Week 01, Reproducible research: In this unit we explore the basics of statistical computing. We look at examples and benefits of plain text formats for data, code, and analyses.

install R, RStudio and user-contributed packages

understand basic version control using the web-based GitHub GUI

create CSV files and read them into R

basic techniques for accessing variables in data objects

selecting variable names

constructing a data dictionary

Week 02, Variable types and numeric summaries: We begin our study of tabular data, with observations stored in rows and variables stored in columns. We start by describing the types of data that can be stored. Next, methods for summarizing and graphing numeric and categorical data are developed.

describe and compute means and medians (using R and by hand)

describe and compute quantiles (using R and, for simple cases, by hand)

describe and compute the standard deviation

differentiate between numeric and categorical data

treating numeric data as categorical data

create categorical variables by grouping numeric data

Week 03, The Grammar of Graphics: Building off of our basic plots, here we describe a self-contained system for the creation of statistical graphics. Putting the topics together allow for the construction of arbitrarily complex visualizations of data.

the basic data aesthetics (x, y, label, color, size, and alpha)

the **ggplot2** syntax

mapping variables to aesthetics

setting fixed aesthetics

constructing layers of points, lines

Week 04, Advanced Graphics: Here we build off of the grammar of graphics to include other layer types, manual annotations, and building professional graphics.

describe scales and themes in the grammar of graphics

use manual annotations to give context to graphics

make use of multiple datasets within a single plot

faceting by categorical variables

Week 05, Filtering and Summarizing Data: Often, data is given or found in a different format than is required for an analysis. In this unit, we study techniques for filtering and restructuring data. One particular focus is the study of how to change the *level of analysis* of a data set, such as taking data originally about individuals and turning it into a dataset to study cities or counties.

syntax of the filter command

boolean variables

binary operators: “and”, “or”

binary operators: greater than, less than

set containment

random subsets

syntax of the summarize command

counting grouped data

Week 06, Statistical Inference: Given a sample of data taken from a larger population, we can use models to estimate how well the sample resembles the entire population. This unit covers an introduction to these techniques, known as statistical inference. The same techniques can be used to study the outcome of random processes.

sample and population statistics

independence

standard errors

confidence intervals

t-tests

Week 07, Data Analysis and Review: A review of the prior weeks and applications to a new dataset. First midterm on Thursday.

Week 08, Communicating Statistical Results: In this unit we cover how to construct arguments using evidence derived from data.

deductive versus inductive reasoning

understanding audience (technical vs. general)

describe exploratory work and hypothesis / thesis generation describe inferential statistics and hypothesis validation

include graphical annotations

Spring Break

Week 09, Normalized Data: In this unit, we will explore and implement best practices for collecting and organizing data.

determining table variables

specifying variable types

consistency standards

ISO date and time standards (ISO 8601)

standards for location data: country codes (ISO 3166), languages (ISO 639), currency (ISO 4217), and US FIPS codes

the tidy data model: rows, columns, and tables

Week 10, Joining Relational Data: We also build off of the last week's material by exploring concepts of data manipulation and data collection to describe methods for working simultaneously with many tables linked together by common keys.

primary keys

foreign keys

composite keys

inner and outer joins

filtering joins

Week 11, Working with strings: Here, we study techniques for using manipulating data stored as strings. We use the stringi library in R to apply functions using fixed strings as well as a standard for describing patterns called regular expressions. You will become familiar with the following tasks and concepts:

detecting substrings

extracting substrings

removing substrings

counting substrings

describing repeating patterns

denoting letters, numbers, and word boundaries

anchoring regular expressions

the UTF-8 encoding

ICU and ISO-639

Week 12, Text mining: This unit builds off of the basic string processing tasks to study textual corpora. You will become familiar with the following concepts and comfortable applying them to new, raw textual data:

tokenization

term-frequency matrices

lemmatization

part-of-speech tags

dependencies

named entities

Week 13, Ethical Guidelines for Statistical Practice: We discuss ethical issues surrounding the practice of data analysis. These include:

data ownership

informed consent and IRB

data privacy and anonymity

open data

p-hacking / data dredging

problematic variables, discrimination, and proxies

algorithmic decision making

confirmation bias

Week 14, Case Study and Review: A review of the prior weeks and applications to a new dataset. Second midterm on Thursday.