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 G13

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

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 week of class.

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. Your solutions must be uploaded to your GitHub page.

Quizzes:

There will be in-class quizzes on most Tuesdays (weeks 2-13). These will usually be closed notes and serve to verify that you are keeping up with the material. 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.

Midterms:

There will be two in-class midterms, scheduled on the following two class meetings:

- 2018-02-22 (Week 6)
- 2018-04-12 (Week 12)

Each exam will expect you to have understood all of the material up to the point, though the focus will be on understanding of new material.

Projects:

You will also complete three data-oriented projects throughout the semester. These are short written documents that mix code, graphics, and prose to provide a comprehensive analysis of a data set. These must also be uploaded to GitHub.

Participation, Attendance and Late Policy:

You are expected to submit work on-time. Late projects will not be considered for a grade of Honors and will not be accepted in any form after a grace-period of 24 hours. You are expected to both attend and participate in class meetings. Excessive absences will result in receiving a grade of "V" for the semester.

There are no make-up exams or quizzes. In rare instances where students have a valid excuse for missing an exam, I will meet with the student to establish a fair process for determining their final grade (see below).

Final Grades:

Traditional final grades are given by assigning a numeric score to every piece of work done in the course, computing a weighted average of these scores, and then converting this average to a letter grade using a either a fixed or curved set of cut-offs. I believe that this approach yields a poor indicator of student performance and does not do a good job of incentivizing learning.

As an alternative, I grade homework (labs) and quizzes on a Pass/Fail basis, reports on a Honors/Pass/Fail basis, and simply report percentages for the exams. These scores, together with your attendance and course engagement, yield a final grade based on the following criteria. Pluses and minuses are used for students who fall between the categories.

	A	**B**	**C**
Participation	actively present and engaged in almost all classes (miss at most 2)	actively present and engaged in most all classes (miss at most 3)	actively present and engaged in most classes (miss at most 5)
Labs	passing grade on nearly all labs (missing no more than 2)	passing grade on most labs (missing no more than 4)	passing grade on the majority of labs (missing no more than 6)
Quizzes	pass at least 8 quizzes	pass at least 6 quizzes	pass at least 4 quizzes
Reports	at least two honors and no failing grades	pass all three data analyses	pass at least two data analyses
Exams	score above 85% on both exams	score above 85% on both exams	score above 85% on both exams

You are allowed one 'free pass' on these requirements; for example, you can get a B if you fail one data analysis by fulfill all other requirements. Note that failing to meet the requirements for a B will result in a grade no higher than a B; that is, you cannot generally make-up for performance in one area of the course with good performance in another. Failure to meet the requirements for a C may result in a failing grade for the term.

Learning Objectives:

Each week is centered around specific learning objectives, which are described below. These objectives will be tested in the weekly quizzes and exams. The topics below serve only to give a general idea about the direction of the course; the exact topics, pace, and coverage may change over the course of the semester.

Week 01, Reproducible computing: 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

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

Week 07, 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 08, Normalized Data: In this unit, we will explore and implement best practices for collecting and organizing data.

determining table variables

selecting variable names

constructing a data dictionary

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 09, 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 10, Messy Data: Here

Week 11, Case Study and Review: A review of the prior weeks and applications to a new dataset. Second midterm on Thursday.

Week 12, Communicating Statistical Evidence: 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

Week 13, 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 14, 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 matricies

lemmatization

part-of-speech tags

dependencies

named entities