CWRU DSCI351-451: Week 07a MidTermReview

Profs: R. H. French, L. S. Bruckman, P. Leu, K. Davis, S. Cirlos

TAs: W. Oltjen, K. Hernandez, M. Li, M. Li, D. Colvin

11 October, 2022

Contents

7.1.1.1 Class Readings, Assignments, Syllabus Topics

7.1.1.1.1 Reading, Lab Exercises, SemProjects

- Readings:
  - For today: OIS 6.1, 6.2
  - For next class:
- Laboratory Exercises:
  - LE4 : is due Thursday October 20th
  - LE :
- Office Hours: (Class Canvas Calendar for Zoom Link)
  - Wednesday @ 4:00 PM to 5:00 PM, Will Oltjen
  - Saturday @ 3:00 PM to 4:00 PM, Kristen Hernandez
  - **Office Hours are on Zoom, and recorded**
- Semester Projects
  - DSCI 451 Students Biweekly Update 1 Due
  - DSCI 451 Students
    - Next Report Out #2 is Due Friday October 28th
  - All DSCI 351/351M/451 Students:
    - **Peer Grading of Report Out #1 is Due October 11th, 2022**
  - Exams
    - MidTerm: Tuesday October 18th, in class or remote, 11:30 - 12:45 PM
    - Final: Monday December 19, 2022, 12:00PM - 3:00PM, Nord 356 or remote

7.1.1.1.2 Textbooks

- Introduction to R and Data Science
  - For R, Coding, Inferential Statistics
    - Peng: R Programming for Data Science
    - Peng: Exploratory Data Analysis with R

Textbooks for this class

- OIS = Diez, Barr, Çetinkaya-Runde: Open Intro Stat v4
- R4DS = Wickham, Grolemund: R for Data Science

Textbooks for DSCI353/353M/453, And in your Repo now

- ISLR = James, Witten, Hastie, Tibshirani: Intro to Statistical Learning with R
- ESL = Trevor Hastie, Tibshirani, Friedman: Elements of Statistical Learning
- DLwR = Chollet, Allaire: Deep Learning with R

Magazine Articles about Deep Learning

- DL1 to DL6 are "Deep Learning" articles in 3-readings/2-articles/

7.1.1.1.3 Tidyverse Cheatsheets, Functions and Reading Your Code

- Look at the Tidyverse Cheatsheet
  - **Tidyverse For Beginners Cheatsheet**
    - In the Git/20s-dsci353-353m-453-prof/3-readings/3-CheatSheets/ folder
  - **Data Wrangling with dplyr and tidyr Cheatsheet**

  Tidyverse Functions & Conventions

  - The pipe operator `%>%`
  - Use `dplyr::filter()` to subset data row-wise.
  - Use `dplyr::arrange()` to sort the observations in a data frame
  - Use `dplyr::mutate()` to update or create new columns of a data frame
  - Use `dplyr::summarize()` to turn many observations into a single data point
  - Use `dplyr::arrange()` to change the ordering of the rows of a data frame
  - Use `dplyr::select()` to choose variables from a tibble,
    - keeps only variables you mention
  - Use `dplyr::rename()` keeps all the variables and renames variables
    - rename(iris, petal_length = Petal.Length)
  - These can be combined using `dplyr::group_by()`
    - which lets you perform operations "by group".
  - The `%in%` matches conditions provided by a vector using the c() function
  - The **forcats** package has tidyverse functions
    - for factors (categorical variables)
  - The **readr** package has tidyverse functions
    - to read_...., melt_... col_..., parse_... data and objects

Reading Your Code: Whenever you see

- The assignment operator `<-`, think **"gets"**
- The pipe operator, `%>%`, think **"then"**

7.1.1.1.4 Syllabus

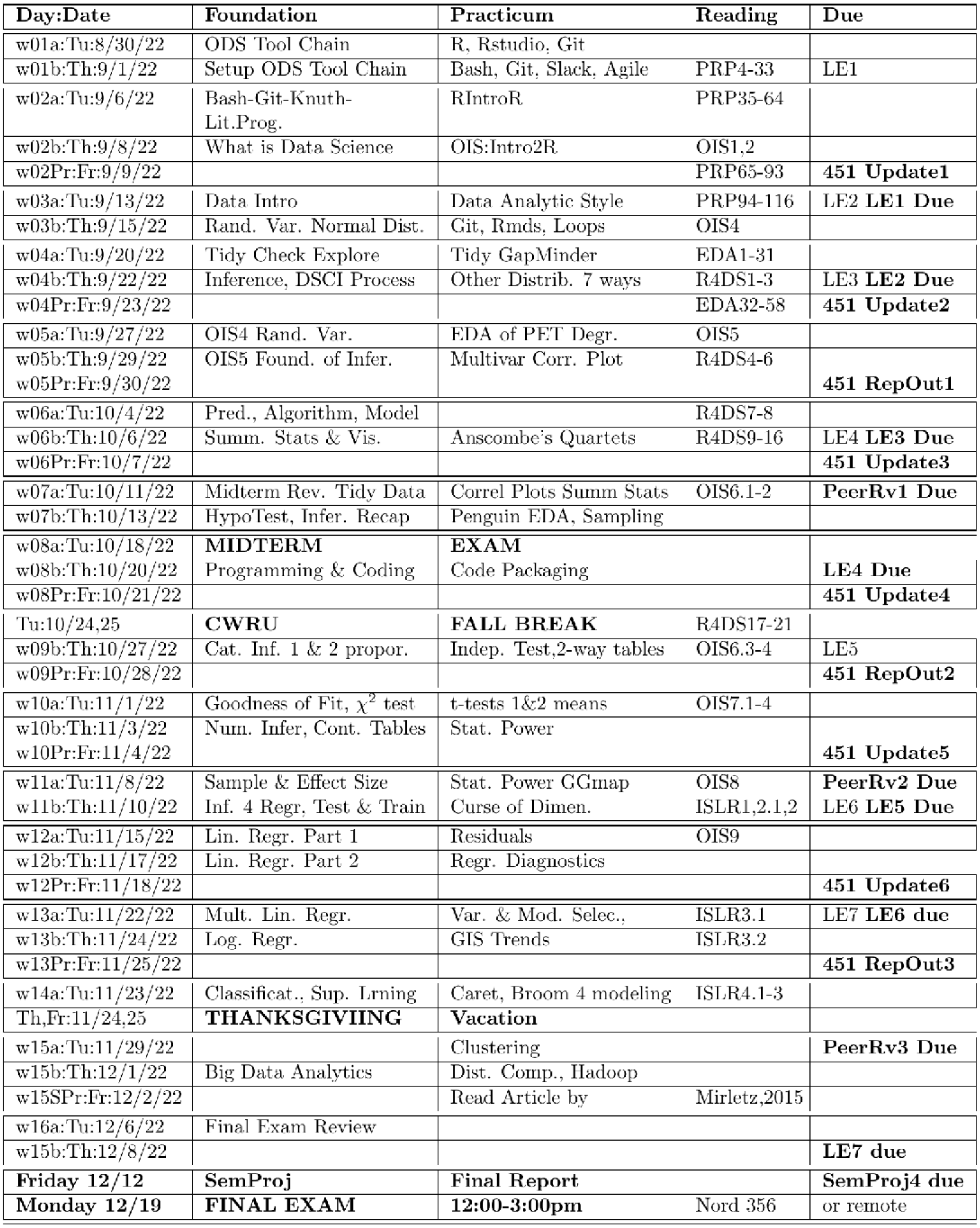
```r
options("digits" = 5)
options("digits.secs" = 3)
library(learningr)
library(tidyverse)
```

7.1.1.1.5 setup for r-code chunks

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.5
## v tibble  3.1.7     v dplyr   1.0.10
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.3     v forcats 0.5.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```



Day:Date	Foundation	Practicum	Reading	Due
w01a:Tu:8/30/22	ODS Tool Chain	R, Rstudio, Git		
w01b:Th:9/1/22	Setup ODS Tool Chain	Bash, Git, Slack, Agile	PRP4-33	LE1
w02a:Tu:9/6/22	Bash-Git-Knuth-Lit.Prog.	RIntroR	PRP35-64	
w02b:Th:9/8/22	What is Data Science	OIS:Intro2R	OIS1,2	
w02Pr:Fr:9/9/22			PRP65-93	**451 Update1**
w03a:Tu:9/13/22	Data Intro	Data Analytic Style	PRP94-116	LE2 **LE1 Due**
w03b:Th:9/15/22	Rand. Var. Normal Dist.	Git, Rmds, Loops	OIS4	
w04a:Tu:9/20/22	Tidy Check Explore	Tidy GapMinder	EDA1-31	
w04b:Th:9/22/22	Inference, DSCI Process	Other Distrib. 7 ways	R4DS1-3	LE3 **LE2 Due**
w04Pr:Fr:9/23/22			EDA32-58	**451 Update2**
w05a:Tu:9/27/22	OIS4 Rand. Var.	EDA of PET Degr.	OIS5	
w05b:Th:9/29/22	OIS5 Found. of Infer.	Multivar Corr. Plot	R4DS4-6	
w05Pr:Fr:9/30/22				**451 RepOut1**
w06a:Tu:10/4/22	Pred., Algorithm, Model		R4DS7-8	
w06b:Th:10/6/22	Summ. Stats & Vis.	Anscombe's Quartets	R4DS9-16	LE4 **LE3 Due**
w06Pr:Fr:10/7/22				**451 Update3**
w07a:Tu:10/11/22	Midterm Rev. Tidy Data	Correl Plots Summ Stats	OIS6.1-2	**PeerRv1 Due**
w07b:Th:10/13/22	HypoTest, Infer. Recap	Penguin EDA, Sampling		
w08a:Tu:10/18/22	**MIDTERM**	**EXAM**		
w08b:Th:10/20/22	Programming & Coding	Code Packaging		**LE4 Due**
w08Pr:Fr:10/21/22				**451 Update4**
Tu:10/24,25	**CWRU**	**FALL BREAK**	R4DS17-21	
w09b:Th:10/27/22	Cat. Inf. 1 & 2 propor.	Indep. Test,2-way tables	OIS6.3-4	LE5
w09Pr:Fr:10/28/22				**451 RepOut2**
w10a:Tu:11/1/22	Goodness of Fit, χ^2 test	t-tests 1&2 means	OIS7.1-4	
w10b:Th:11/3/22	Num. Infer, Cont. Tables	Stat. Power		
w10Pr:Fr:11/4/22				**451 Update5**
w11a:Tu:11/8/22	Sample & Effect Size	Stat. Power GGmap	OIS8	**PeerRv2 Due**
w11b:Th:11/10/22	Inf. 4 Regr, Test & Train	Curse of Dimen.	ISLR1,2.1,2	LE6 **LE5 Due**
w12a:Tu:11/15/22	Lin. Regr. Part 1	Residuals	OIS9	
w12b:Th:11/17/22	Lin. Regr. Part 2	Regr. Diagnostics		
w12Pr:Fr:11/18/22				**451 Update6**
w13a:Tu:11/22/22	Mult. Lin. Regr.	Var. & Mod. Selec.,	ISLR3.1	LE7 **LE6 due**
w13b:Th:11/24/22	Log. Regr.	GIS Trends	ISLR3.2	
w13Pr:Fr:11/25/22				**451 RepOut3**
w14a:Tu:11/23/22	Classificat., Sup. Lrning	Caret, Broom 4 modeling	ISLR4.1-3	
Th,Fr:11/24,25	**THANKSGIVIING**	**Vacation**		
w15a:Tu:11/29/22		Clustering		**PeerRv3 Due**
w15b:Th:12/1/22	Big Data Analytics	Dist. Comp., Hadoop		
w15SPr:Fr:12/2/22		Read Article by	Mirletz,2015	
w16a:Tu:12/6/22	Final Exam Review			
w15b:Th:12/8/22				**LE7 due**
Friday 12/12	**SemProj**	**Final Report**		**SemProj4 due**
Monday 12/19	**FINAL EXAM**	**12:00-3:00pm**	Nord 356	or remote

Figure 1: DSCI351-351M-451 Syllabus

7.1.1.2 Midterm

- Testing Concepts, OpenIntro Stats, and R for Data Science
- Your Data Science Tool Chain
- Open and Reproducible Science
- Steps in Data Analysis
- Done as Rmd and Rscripts

7.1.1.2.1 Today: Get your class repo in good shape

- `git status`
- `git pull`
- `git status`
- `git add --all :/`
- `git status`
- `git commit -m 'getting my repo synchronized with bitbucket website'`
- `git push`

7.1.1.2.2 Make Sure You Have OnDemand/Markov and MyApps/ODS Desktop

- Setup
- With a copy of your personal course repo
- Location for you Git Repo
  - Markov: `/mnt/pan/courses/dsci351-451/CaseID/Git/`
  - ODS Desktop: `H:\Git\`

7.1.1.2.3 Since we are Synch/Asynch/Remote

- When you start exam, put the time in the top of your exam
  - I'll have a spot to put that
- Then we will see you time for submission to assignment page
- You have a total of 1.5 hours to do the exam
- If you need special arrangements,
  - Tell Raymond and I in a direct message in DSCI Class Slack

CWRU's Academic Integrity Policy

- https://bulletin.case.edu/undergraduatestudies/academicintegrity/

7.1.1.2.4 Midterm is open book / open resource

- The midterm will be given as an Rmd
- You will work in the Rmd file
- Writing and doing Rcode chunks
- You have the resources of
  - Your repository
  - R Help
  - Other online resources
- Open Data Science Approach
  - What can you accomplish
  - Using all available resources

7.1.1.2.5 Midterm Covers

- [Roger Peng's R Programming, EDA with R]
- **R4DS Chapters 1 - 16**
- **OIS Chapters 1,2,4,5** i.e. Through Foundations of Inference

    - Foundations of Inference (OIS-5)
    - Slides from OIS are in 3-readings\1-Textbooks\2-OpIntStats-slides
- **Jeff Leek's Structure of a Data Analysis**
    - In class notes: 2108-351-351m-451-w03a-p1-DataAnalyticStyle
    - Also Chap. 14 of Leek's The Elements of Data Analytic Style
        * In 3-readings\1-Textbooks

7.1.1.2.6 Midterm Does Not Cover OIS Chapters 6 and beyond

- Inference for Numerical Data (OIS-6)
- Inference for Categorical Data (OIS-7)

7.1.1.2.7 Midterm doesn't cover linear modeling

- Such as Regression
- Or Classification

7.1.1.2.8 Topics Covered In Class

- both Foundations and Practicum topics

7.1.1.3 Midterm Concepts

- e. g. Open Data Science, Data Analysis, EDA, Visualization, Inference
    - Git, Rstudio, R, R packages
    - Graphics and Visualization: Base and GGPlot2
    - Data Assembly, Cleaning
    - Exploratory Data Analysis
    - Tidyverse: Pipes, dplyr, mutate etc.
    - Study Design
    - Distributions, and Central Limit Theorem
    - Sampling and Populations
    - Standard Errors, Confidence Intervals
    - Hypothesis Testing
    - Summary Statistics & Visualization
    - Multivariate pairwise correlation plots
    - Other topics

7.1.1.4 Data Science Tool Chain

7.1.1.4.1 Openness:

- https://en.wikipedia.org/wiki/Openness
- Free and open source software (FOSS)
    - Open source code and programs
    - https://en.wikipedia.org/wiki/Open-source_model
- Open datasets
- https://en.wikipedia.org/wiki/Open_data
- Open Access
    - https://en.wikipedia.org/wiki/Open_access

7.1.1.4.2 R statistics programming language

- > 20,000 packages,

Python

- Also a good statistical environment
- not as well developed for stats
- but better are substantial number crunching

There are many other stats softwares and languages

- SPSS, SAS, STATA,
  - But these are not useful for automated analysis

7.1.1.4.3 But Excel, or mousey/mousey programs are not for data science

- Can not record the sequential processing
  - i.e. the script of your analysis
- don't lead to reproducible and open science
- can't distribute code, data and analysis and report

7.1.1.4.4 IDE (Integrated Development Environment)

- Comfortable environment for getting going
- Rstudio for R,
- PyCharm or Spyder or Eclipse with PyDev for Python

7.1.1.4.5 Yet everything can be done at the command line

- This enables automation
- And large scale analysis
- Using scripting (bash scripting)
- Simple automation

7.1.1.4.6 Git Repositories for content versioning

- Can pursue branches and revert to earlier versions
- Enables collaboration
- Robust code review
- Fork and develop in a community
- IDEs support Git Versioning

7.1.1.4.7 Markdown languages

- Enable integrated reports, code, data in repositories
- RMarkdown2 for R
- iPython Notebooks for Python
- And Report can autoupdate with a simple re-compile

Direction towards interactive data science

7.1.1.5 Peng's R Programming (PRP) and Exploratory Data Analysis (EDA)

7.1.1.5.1 Using R as a calculator

- Mathematical operations and vectors
- Assigning variables
- Special numbers
- Logical vectors

7.1.1.5.2 Inspecting variables and your workspace

- Classes
- Different types of numbers
- Other common classes
- Checking and changing classes
- Examining variables
- The workspace

7.1.1.5.3 Vectors, matrices and Arrays, List & Dataframes

- Vectors
- Matrices & Arrays
- Lists
- Data Frames
- – Creating Data Frames
- – Indexing Data Frames
- – Basic Data Frame Manipulation

7.1.1.5.4 Environments & Functions

- Environments
- Functions
- – Creating and Calling Functions
- – Passing Functions to and from Other Functions
- – Variable Scope

7.1.1.5.5 Strings & Factors

- Strings
- – Constructing and Printing Strings
- – Formatting Numbers
- – Special Characters
- – Changing Case
- – Extracting Substrings
- – Splitting Strings
- – File Paths
- Factors
- – Creating Factors
- – Changing Factor Levels
- – Dropping Factor Levels
- – Ordered Factors
- – Converting Continuous Variables to Categorical

- – Converting Categorical Variables to Continuous
- – Generating Factor Levels
- – Combining Factors

7.1.1.5.6 Getting Data

- Built-in Datasets
- Reading Text Files
- – CSV and Tab-Delimited Files
- – Unstructured Text Files
- – XML and HTML Files
- – JSON and YAML Files
- Reading Binary Files
- Web Data
- – Sites with an API
- – Scraping Web Pages

7.1.1.5.7 Cleaning and Transforming (Tidying)

- Cleaning Strings
- Manipulating Data Frames
- – Adding and Replacing Columns
- – Dealing with Missing Values
- – Converting Between Wide and Long Form
- – Using SQL
- Sorting

7.1.1.5.8 Exploring and Visualizing (EDA)

- Summary Statistics
- The Three Plotting Systems
- – Take 1: base Graphics
- – (We Ignore)Take 2: lattice Graphics
- – Take 3: ggplot2 Graphics
- Scatterplots
- Line Plots
- Histograms
- Box Plots
- Bar Charts
- Other Plotting Packages and Systems

7.1.1.5.9 So in DSCI

- Your learning coding
- statistical concepts, tools, and approaches
- open data science methods
- open collaboration and learning approaches

7.1.1.6 R for Data Science (R4DS)

- Tidyverse functions
  - Tidy dataframes

7.1.1.6.1 Writing R scripts and the R console

- Moving around RStudio

- Features of the R console
- Features of the source editor

7.1.1.6.2 Viewing and Plotting Data

- Object Browser
- Plotting
- Plotting with Manipulate Package

7.1.1.6.3 Managing R Projects

- R Projects
- Version Control with Git

7.1.1.6.4 Generating Reports (Open Data Science)

- R markdown
- Code Chunks
- LaTeX

7.1.1.6.5 Literate Programming (or Open/Reproducible Data Science)

- Finally, we note that the interweaving of code and text
  - (often referred to as literate programming) may serve two purposes.
  - The first is to generate a data analysis report
    - by executing code to produce the result.
  - The second is to document the code itself, for example,
    - by describing the purpose of a function and all its arguments.

The latter purpose will be discussed

- with the Roxygen2 package for code documentation.

7.1.1.7 What is a Data Analysis

7.1.1.7.1 Steps in a Data Analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data (Open/Available Data first for pilot study)

- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

7.1.1.8 Open Intro Statistics

7.1.1.8.1 Open Intro Stats: OI-1 Intro to Data, OI-2 Summarizing Data

- Data basics
- Overview of data collection principles
- Observational studies and sampling strategies
- Experiments
- Examining numerical data
- Considering categorical data

7.1.1.8.2 OI-4 Distributions of Random Variables

- Normal distribution
- Evaluating the normal approximation
- Geometric distribution
- Binomial distribution
- BUT NOT the Poisson Distribution

7.1.1.8.3 OI-5 Foundations of Inference

- Variability in estimates
- Confidence intervals
- Hypothesis intervals
- Examining the central limit theorem
- Inference for other estimators
- Sample size and power
- Statistical vs. practical significance

7.1.1.8.4 So Things to know

- Z values (# of sd's away from mean)
- zstar values
- normal probability plots
- How to form a hypothesis for hypothesis testing
- p values
- Type I and II errors
- alpha and beta values
- census vs. sampling
- observational studies, controlled studies
- prospective studies and retrospective studies
- IQRs interquartile ranks
- SE (standard error of an estimate)
- SE of the sample mean
- population values vs. point estimates: mu vs xbar
- Confidence Intervals, 95% CIs

7.1.1.8.5 Conditions for $\bar{x}$ ("xbar") being nearly normal and Standard Error (SE) being accurate

- Important conditions to help ensure the sampling distribution of x
  - is nearly normal and
  - the estimate of SE sufficiently accurate:

Requires

- The sample observations are independent.
- The sample size is large: n = 30 (or n > 30)
  - is a good rule of thumb.
- The distribution of sample observations
  - is not strongly skewed.

Additionally, the larger the sample size

- the more lenient we can be with the sample's skew.

7.1.1.9 THE FOLLOWING TOPICS NOT ON MIDTERM:

- Probability, OIS Chapter 3
- Numerical Inference, OIS Chapter 6
- Categorical Inference, OIS Chapter 7

7.1.1.10 During the MidTerm Exam Class Time

- We will have the class Zoom session On
- So that you can ask any questions

I will `git push` the midterm exam at 11:30 to my Prof Repo

- You sync Prof repo with your class repo - up on Bitbucket website
- The `git pull` your class repo
- Change …NAME.Rmd to your …caseID.Rmd

7.1.1.10.1 And Turn In the exam by 12:45 or 1:00 PM

- Compile to pdf
  - And submit your .Rmd and .pdf
- If you have a problem compiling
  - Then compile to .html
  - And submit .Rmd and .html