CWRU DSCI351-451: MidTermReview

04 October, 2018

Contents

7.1.0.1 Reading, Homeworks, Projects, SemProjects . . . 1
7.2 Readings: . . . 2
7.2.0.1 Syllabus . . . 2
7.2.0.2 setup for r-code chunks . . . 2
7.2.0.3 Midterm . . . 2
7.2.0.3.1 Midterm is open book / open resource . . . 2
7.2.0.3.2 Midterm Does Not Cover Foundations of Inference . . . 4
7.2.0.3.3 Topics Covered In Class . . . 4
7.2.0.4 Midterm Concepts e. g. Open Data Science, Data Analysis, EDA, Visualiation 4
7.2.0.5 R statistics programming language . . . 4
7.2.0.5.1 But Excel, or mousey/mousey programs are not for data science . . 4
7.2.0.5.2 IDE (Integrated Development Environment) . . . 4
7.2.0.5.3 Yet everything can be done at command line . . . 5
7.2.0.5.4 Git Repositories for content versioning . . . 5
7.2.0.5.5 Markdown languages . . . 5
7.2.0.6 Peng's R Programming (PRP) and Exploratory Dati Analysis (EDA) . . . 5
7.2.0.6.1 Using R as a calculator . . . 5
7.2.0.6.2 Inspecting variables and your workspace . . . 5
7.2.0.6.3 Vectors, matrices and Arrays, List & Dataframes . . . 5
7.2.0.6.4 Environments & Functions . . . 6
7.2.0.6.5 Strings & Factors . . . 6
7.2.0.6.6 Getting Data . . . 6
7.2.0.6.7 Cleaning and Transforming (Tidying) . . . 7
7.2.0.6.8 Exploring and Visualizing (EDA) . . . 7
7.2.0.7 So in DSCI . . . 7
7.2.0.8 R for Data Science (R4DS) . . . 7
7.2.0.8.1 Writing R scripts and the R console . . . 7
7.2.0.8.2 Viewing and Plotting Data . . . 7
7.2.0.8.3 Managing R Projects . . . 8
7.2.0.8.4 Generating Reports (Open Data Science) . . . 8
7.2.0.8.5 Literate Programming (or Open/Reproducible Data Science) . . . 8
7.2.0.9 What is a Data Analysis . . . 8
7.2.0.9.1 Steps in a Data Analysis . . . 8
7.2.0.9.2 Open Intro Stats: OI-1 Intro to Data . . . 8
7.2.0.10 THE FOLLOWING TOPICS NOT ON MIDTERM: Inferential Statistics . . 8
7.2.0.10.1 OI-3 Distributions of Random Variables . . . 9
7.2.0.10.2 OI-4 Foundations of Inference (Not on Exam) . . . 9
7.2.0.10.3 So Things to know (Not on Exam) . . . 9
7.2.0.10.4 Conditions for xbar being nearly normal and SE being accurate (Not on Exam) . . . 9

7.1.0.1 Reading, Homeworks, Projects, SemProjects

- Homework:
  - HW 4 release on Thursday October 12th
  - HW 4 Due Tuesday October 17 before class

-

7.2 Readings:

- 451 SemProjects:
  -

7.2.0.1 Syllabus

7.2.0.2 setup for r-code chunks

- rmarkdown::render('1502-w06b-f-FrenchDSCI351-451-numerical-inference.Rmd', 'all')

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

```
## -- Attaching packages ---------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

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

7.2.0.3 Midterm

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

7.2.0.3.1 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

Day:Date	Foundation	Practicum	Reading	Due
w1a:Tu:8/28/18	ODS Tool Chain	R, Rstudio, Git		
w1b:Th:8/30/18	Setup ODS Tool Chain	Bash, Git, Twitter	PRP4-33	HW1
w2a:Tu:9/4/18	What is Data Science	OIS:Intro2R	PRP35-64	**HW1 Due**
w2b:Th:9/6/18	Data Analytic Style, Git	451SempProj, Git	PRP65-93, OI1-1.9	HW2
w3a:Tu:9/11/18*	Struct. of Data Analysis	ISLR:Intro2R, Loops	PRP94-116, OIS3	**HW2 Due**
w3b:Th:9/13/18*	OIS3 Intro to Data	GapMinder, Dplyr, Magrittr		
w4a:Tu:9/18/18	OIS3, Intro2Data part 2, Data	EDA: PET Degr.	EDA1-31	Proj1
w4b:Th:9/20/18	Hypothesis Testing	GGPlot2 Tutorial	EDA32-58	HW3
w5a:Tu:9/25/18	Distributions	SemProj RepOut1	R4DS1-3	**HW3 Due**
w5b:Th:9/27/18	Wickham DSCI in Tidyverse	SemProj RepOut1	R4DS4-6	**SemProj1,**
w6a:Tu:10/2/18	OIS Found. of Inference	Inference	R4DS7-8	**Proj1 Due**
w6b:Th:10/4/18		Midterm Review	R4DS9-16 Wrangle	
w7a:Tu:10/9/18*	Summ. Stats & Vis.	Data Wrangling		
w7b:Th:10/11/18*	**MIDTERM EXAM**			HW4
w8a:Tu:10/16/18	Numerical Inference	Tidy Check Explore	OIS4	**HW4 Due**
w8b:Th:10/18/18	Algorithms, Models	Pairwise Corr. Plots	OIS5.1-4	Proj 2, HW5
Tu:10/23	**CWRU FALL BREAK**		R4DS17-21 Program	
w9b:Th:10/25/18	Categorical Infer	Predictive Analytics	OIS6.1,2	
w10a:Tu:10/30/18	SemProj	SemProj	OIS7	**SemProj2 HW5 Du**
w10b:Th:11/1/18	Lin. Regr.	Lin. Regr.	OIS8	**Proj.2 due**
w11a:Tu:11/6/18	Inf. for Regression	Curse of Dim.	OIS8	Proj 3
w11b:Th:11/8/18	Model Accuracy	Training Testing	ISLR3	HW6
w12a:Tu:11/13/18	Multiple Regr.	Mul. Regr. & Pred.	ISLR4	**HW6 due**
w12b:Th:11/15/18	Classification		ISLR6	
w13a:Tu:11/20/18	Classification	Clustering	ISLR5	**Proj 3 due**
Th:11/22/18	**THANKSGIVING**			Proj 4
w14a:Tu:11/27/18	Big Data	Hadoop		
w14b:Th:11/29/18	InfoSec	VerisDB		**SemProj3**
w15a:Tu:12/4/18	SemProj ReportOut3			
w15b:Th:12/6/18	SemProj ReportOut3			**Proj4**
	FINAL EXAM	**Monday12/17, 12:00-3:00pm**	Olin 313	**SemProj4 due**

Figure 1: DSCI351/451 Syllabus

- – Using all available resources

7.2.0.3.2 Midterm Does Not Cover Foundations of Inference

- Foundations of Inference (OIS-4)
- Inference for Numerical Data (OIS-5)
- Inference for Categorical Data (OIS-6)

7.2.0.3.3 Topics Covered In Class

- both Foundations and Practicum topics

7.2.0.4 Midterm Concepts e. g. Open Data Science, Data Analysis, EDA, Visualiation

- Git, Rstudio, R, R packages
- Graphics, Base and GGPlot2
- Data Assembly, Cleaning
- Exploratory Data Analysis
- Tidyverse: Pipes, dplyr, mutate etc.
- Study Design
- Sampling and Populations
- Other topics

Data Science Tool Chain

7.2.0.5 R statistics programming language

- > 8000 packages, free and open source software (FOSS)

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

7.2.0.5.1 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

7.2.0.5.2 IDE (Integrated Development Environment)

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

7.2.0.5.3 Yet everything can be done at command line

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

7.2.0.5.4 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

7.2.0.5.5 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

7.2.0.6 Peng's R Programming (PRP) and Exploratory Dati Analysis (EDA)

7.2.0.6.1 Using R as a calculator

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

7.2.0.6.2 Inspecting variables and your workspace

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

7.2.0.6.3 Vectors, matrices and Arrays, List & Dataframes

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

7.2.0.6.4 Environments & Functions

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

7.2.0.6.5 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

7.2.0.6.6 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

7.2.0.6.7 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

7.2.0.6.8 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

7.2.0.7 So in DSCI

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

7.2.0.8 R for Data Science (R4DS)

7.2.0.8.1 Writing R scripts and the R console

- – Moving around RStudio
  - – Features of the R console
  - – Features of the source editor

7.2.0.8.2 Viewing and Plotting Data

- Object Browser
- Plotting
- Plotting with Manipulate Package

7.2.0.8.3 Managing R Projects

- R Projects
- Version Control with Git

7.2.0.8.4 Generating Reports (Open Data Science)

- R markdown
- Code Chunks
- LaTeX

7.2.0.8.5 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.

7.2.0.9 What is a Data Analysis

7.2.0.9.1 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

7.2.0.9.2 Open Intro Stats: OI-1 Intro to Data

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

7.2.0.10 THE FOLLOWING TOPICS NOT ON MIDTERM: Inferential Statistics

7.2.0.10.1 OI-3 Distributions of Random Variables

- Normal distribution
- Evaluating the normal approximation
- Geometric distribution
- Binomial distribution

7.2.0.10.2 OI-4 Foundations of Inference (Not on Exam)

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

7.2.0.10.3 So Things to know (Not on Exam)

- 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

7.2.0.10.4 Conditions for xbar being nearly normal and SE being accurate (Not on Exam)

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

- 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.