INT3042: Introduction to Social Science Data Analysis

Yonsei University, Wonju

Fall 2012

Version: 23 August 2012

Instructor: Dr. Christopher Gandrud

- Email: gandrud@yonsei.ac.kr
- Website: http://christophergandrud.github.com/

Office Hours: 15:00-17:00 Wednesday (정208)

- You can also send me an email or come to my office whenever you need to.

Time: 11:00-12:15 Tuesday & Thursday (정346)

Objectives: This course's main objective is to: *learn how to take raw social science data, explore it, and present the results in a useful way*. In this course you will learn all of the basic skills needed to do each of these steps with the statistical language **R**. *Part I* of the course introduces you to both basic data structures and **RStudio** (a program that makes using **R** easier). *Part II* introduces basic data analysis and visualizations techniques. *Part III* covers slightly more advanced statistical tools, primarily linear regression. Finally, in *Part IV* we will apply all of these skills in a pair research project.

As part of achieving this straightforward objective, the course is intended to also do the following:

The course is intended to be *useful*. I hope that the course will be one of the more useful courses you take in university. It is intended to be useful for students who want to go on to do graduate-level *academic research*. It is also intended to be useful for students who want to go directly into the *non-academic public or private sectors* and be able to effectively analyse and present data. I hope it is also useful for you as a *citizen* in that you will be better able to critically read the data that informs many of our daily decisions.

This course will introduce you to a vibrant and growing community of open source data analysts that is pushing the state of the art in data analysis and presentation. (Another

benefit to you as students is that open source software is free.)

This course emphasises the collection and analysis of social science data (particularly political science and economic data at the country-level). However, almost all of the skills you learn in this course can be applied to data in most other areas of study. **R** is widely used not only in the social sciences but also in medicine, biology, physics, and business.

Prerequisites: If you have taken high school algebra, can use Microsoft Excel, and have an interest in experimenting and learning new things then you are qualified to take this course.

Course Materials: The course syllabus, lecture slides, activities, the course wiki and some data sets can be found on the course's main **github** site:

> https://github.com/christophergandrud/Introduction_to_Statistics_and_Data_Analysis_Yonsei

You are reading the course syllabus now. It has the file name `README.md` on **github**.

Software: We use four software programs in this course:

- Microsoft Excel
- Dropbox
- R (version 2.15)
- RStudio (version 0.96)

All of the classroom computers have these programs installed. Please feel free to use your own computer. You can install both **R** and **RStudio** for free (they are open source!) on Windows, Mac, or Linux computers. Just take two steps:

1. Install **R**: http://www.r-project.org/
2. Install **RStudio**: http://rstudio.org/

The course wiki has slightly more detailed installation instructions.

I recommend using your own laptop computer.

Dropbox is an easy to use cloud storage system. You will be able to use it to access all of your files for the course on almost any computer. You can also submit your assignments to me through **Dropbox**. You can download **Dropbox** from https://www.dropbox.com/. You will also need to sign up for a (free) account.

Microsoft Excel is not free, but you probably already have it. If you don't, you can use any other spreadsheet program like Hancell, Numbers, or Open Office. Open Office is open source and can be downloaded for free from http://www.openoffice.org/.

Readings: This is an applied course, so I want you to *do* more than you *consume* (read,

listen to lectures). However, completing the course readings before each class is an important part of (a) being prepared to do the work in class, (b) obtaining a solid foundation in the methods and theory behind what we do in class, and © pointing you to useful reference material.

The main textbook for this class is *OpenIntro Statistics (first edition)*. You can find the PDF at:

> http://www.openintro.org/stat/downloads.php

OpenIntro Statistics is an open source introductory statistics textbook written mostly by David M. Diez, Christopher D. Barr, and Mine Çetinkaya-Rundel.

Most of the other readings for the course will be academic articles or blogposts by academic bloggers. I hope these will introduce you to the very active social science blogging community who are advancing the state of the art in data management and analysis.

I also assign a few chapters from (and recommend for **R** reference):

> Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.
>
> Matloff, Norman. 2011. *The Art of R Programming: A Tour of Statistical Design Software*. San Francisco: No Starch Press.

For more reference on using **R** you might find these resources useful:

- *Overall R reference:* Johnson, Paul, E. 2012. “Rtips. Revivial 2012”. http://pj.freefaculty.org/R/Rtips.html.
- *Generally cutting-edge posts on how to use R:* R-Bloggers. http://www.r-bloggers.com/
- *Interesting blog on how someone at the New York TImes creates graphs:* http://chartsnthings.tumblr.com/

Assessment:

- 10% Class Attendance and Participation
- 40% 5 Short Assignments: Due weeks 3, 5, 7, 9, 11
- 50% Pair Research project (paper and presentation): Due Week 16

How we communicate recently has been changing rapidly. Inaddition to articles and presentations, people have been using new formats such as **Twitter** and blogs to share their work and ideas. To help develop your skills using these formats you will post your

work for some of the short assignments and pair research project on the class blog. You should post both the output of your work and the code you used to create it. The blog is at:

> http://yonseiappliedstats.tumblr.com/

I'll give you more details about how post on the blog in class. Other details about the assignments will also be given in class.

Part I: Introduction to Data Gathering and Management in R

Week 1: Course Introduction and R Introduction

I cover the course objectives, give you some examples of the type of data analysis and visualizations you will learn how to do in the course, and start working with **RStudio**.

Extra:

> **Introduction:** Matloff, Norman. 2011. *The Art of R Programming: A Tour of Statistical Design Software*. San Francisco: No Starch Press.
>
> **Appendix 1: Fundamentals of the R Language:** Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.

Week 2: Types of Data

> **p. 1-8, 26-36:** *OpenIntro Statistics*
>
> **Chapter 2, Dataframes:** Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.

Week 3: Gathering Data

Where can we find quality social science data? How do we get this into **R** so we can start analysing it?

> "Data APIs/Feeds Available as Packages in R". 2011. Cross Validated. http://stats.stackexchange.com/questions/12670/data-apis-feeds-available-as-packages-in-r.

Week 4: Replication!

How to document your work so that others can reproduce it. Why reproducibility is so

important in general and in the social sciences in particular. We will also learn the basics of creating dynamic reports with `knitr` and **Markdown**.

> King, Gary. 1995. “Replication, Replication.” PS: Political Science and Politics 28(3): 444–452.

Part II: Basic Data Analysis and Visualization

Week 5: Descriptive Statistics

What kinds of simple statistical summaries can we use to understand and present our data?

> **p. 9-26:** *OpenIntro Statistics*

Extra:

> **Chapter 3, Central Tendency:** Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.
>
> **Chapter 4: Variance:** Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.

Week 6: Data Visualisation in R

How can we use **R** to create visual summaries of our data? What is `ggplot2` ? What is `googleVis` ?

> Chen, Edwin. 2012. “Quick Introduction to ggplot2” http://blog.echen.me/2012/01/17/quick-introduction-to-ggplot2/
>
> **Chapter 1, Graphical Excellence:** Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press.
>
> **Chapter 4, Data-Ink and Graphical Redesign:** Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press.
>
> **Chapter 9, Aesthetics and Technique in Data Graphical Design:** Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd Edition. Cheshire, Connecticut: Graphics Press.

Extra:

> Donahue, Rafe M. J. 2011. *Fundamental Statistical Concepts in Presenting Data: Principles for Constructing Better Graphics*. Version 2.11.

> http://biostat.mc.vanderbilt.edu/wiki/pub/Main/RafeDonahue/fscipdpfcbg_currentversion.pdf
>
> *If you have access to Adobe Illustrator and want to make commercial journal quality graphics with **R** see:* Sigal, Mathew. 2011. *Make it Pretty:* An Introduction to Graphical Post-Processing with Adobe Illustrator. Presentation to York University Department of Psychology Quantitative Methods Brownbag. http://www.psych.yorku.ca/quantmethods/BrownBag/Sigal-2011-Post-Processing-Handouts.pdf

Week 7: Overview of Statistical Inference

How confident are we that our findings in our samples are actually present in the real world?

> **Chapter 4:** *OpenIntro Statistics*
>
> **Chapter 1, Fundamentals:** Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.

Week 8: Statistical Inference with Large Samples

What basic tools can we use to make inferences when we have large data samples?

> **Chapter 5:** *OpenIntro Statistics*

Extra:

> **Chapter 5, Single Samples:** Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.
>
> **Chapter 6, Two Samples:** Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.

Part III: Introduction to Linear Regression

Week 10: Simple Linear Regression

What is simple linear regression? Why is it so useful for social scientists? How can we use `Zelig` to for regression analysis?

> **Chapter 7:** *OpenIntro Statistics*
>
> **Chapter 8, Regression:** Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.

Extras

> **Chapter 7, Statistical Modelling:** Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.
>
> **p. 305-311:** Imai, Kosuke, Gary King, and Olivia Lau. 2012. *Zelig: Everyone's Favorite Statistical Software.* http://r.iq.harvard.edu/docs/zelig.pdf

Week 11: Multiple Linear Regression

What if we have more than one independent variable?

> **Chapter 8:** *OpenIntro Statistics*

Extra:

> **Chapter 11, Multiple Regression:** Crawley, Michael J. 2005. *Statistics: An Introduction Using R*. Chichester: John Wiley & Sons. Ltd.

Week 12: Presenting Regression results

How can we most effectively present the results from our regression analyses using simulations and graphs? How can we use `Zelig` for regression analysis?

> King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." American Journal of Political Science 44(2): 347–361.

Part IV: Applied Data Analysis: Pair Research Projects

In the final part of the course class time is dedicated to applying what we have learned so far to actually gather, analyse, and present data.

Week 12: Research Question Design and Data Download

Week 13: Data Clean Up and Exploratory Analysis

Week 14: Statistical Analysis and Visualize Results

Week 15: Write Up Results and Prepare Presentations

Week 16: Research Project Presentations