University of Connecticut

School of Business

OPIM 5603

Statistics in Business Analytics

Fall 2019

Section B12

Wednesday 6 – 9 pm

Classroom GBLC 505

OPIM 5603 Fall 2019

Slide 0.2

pronounced **Slavin**

Instructor: Slaven Stricevic

slaven.stricevic@uconn.edu

TAs: Rohit Kavikondala

sai.kavikondala@uconn.edu

Shiyu Ding

shiyu.ding@uconn.edu

Class web page: UConn *HuskyCT*

Public web page:

https://lms.uconn.edu/bbcswebdav/courses/1198-UCONN-OPIM-5603-SECB12-11953/Public

Link to public web page:

tinyurl.com/fall2019stat

Tentative topics

- *R* overview, descriptive measures, data visualization
- Probability and probability distributions
- Sampling distributions
- Confidence intervals
- Hypothesis testing
- Linear regression
- Maximum likelihood estimation
- Generalized linear models

Course Objectives

- Introduction to *R* to provide students familiarity and confidence with the environment to conduct core statistical analysis,
- Design and conduct basic statistical analysis from first principles,
- Explore some advanced topics through add-on packages.

No prior programming experience is assumed!

Only prerequisites are interest and willingness to learn!

Tentative calendar

- The semester lasts 14 (working) weeks: 2 Wednesdays are booked for exams and the remaining 12 are lectures.
- Exam 1 covers the material from the first six lectures and Exam 2 from the next six lectures.
- Exam 1 is scheduled for October 16th. Note: it 'should' take place on week 7, yet it is pushed one week to allow extra study time.
- Exam 2 takes place on December 4th (last week of the semester).

Week	Su	Mo	Tu	We	Th	Fr	Sa
1	Aug 25	Aug 26	Aug 27	Aug 28	Aug 29	Aug 30	Aug 31
2	Sep 1	Sep 2	Sep 3	Sep 4	Sep 5	Sep 6	Sep 7
3	Sep 8	Sep 9	Sep 10	Sep 11	Sep 12	Sep 13	Sep 14
4	Sep 15	Sep 16	Sep 17	Sep 18	Sep 19	Sep 20	Sep 21
5	Sep 22	Sep 23	Sep 24	Sep 25	Sep 26	Sep 27	Sep 28
6	Sep 29	Sep 30	Oct 1	Oct 2	Oct 3	Oct 4	Oct 5
7	Oct 6	Oct 7	Oct 8	Oct 9	Oct 10	Oct 11	Oct 12
8	Oct 13	Oct 14	Oct 15	Oct 16	Oct 17	Oct 18	Oct 19
9	Oct 20	Oct 21	Oct 22	Oct 23	Oct 24	Oct 25	Oct 26
10	Oct 27	Oct 28	Oct 29	Oct 30	Oct 31	Nov 1	Nov 2
11	Nov 3	Nov 4	Nov 5	Nov 6	Nov 7	Nov 8	Nov 9
12	Nov 10	Nov 11	Nov 12	Nov 13	Nov 14	Nov 15	Nov 16
13	Nov 17	Nov 18	Nov 19	Nov 20	Nov 21	Nov 22	Nov 23
Break	Nov 24	Nov 25	Nov 26	Nov 27	Nov 28	Nov 29	Nov 30
14	Dec 1	Dec 2	Dec 3	Dec 4	Dec 5	Dec 6	Dec 7

What is R?

- Popular 'statistical computing and programming' language.
- Used by many institutions of very different kinds.
- Free, open source software.
- Extensible, with many contributors:
  - 2009: "close to 1,600" different packages, various sources,
  - 2015: over 7 thousand packages available from *Comprehensive R Archive Network* (CRAN),
  - presently close to 14.8 thousand packages available from CRAN.

History of R

- 1976: *S* language developed at *AT&T Bell Labs*.
- 1990's: *S* translated to *R* by Ross Ihaka & Robert Gentleman.
- 2000: *R* v.1.0.0 released.
- Thousands of programmers and statisticians contribute to the effort.
- The current released version is 3.6.1 (2019-07-05)

Why R?

- Seamless interaction between manipulating, analyzing and visualizing your data.
- It is a programming language so actions can be repeated thousands of times.
- It is powerful: in all likelihood it is all you will ever need!
- It is free!
- Not just 'open source':
  - Provides full access to algorithms so nothing is hidden.
  - Infinitely expandable, as it provides a forum for people to contribute.
  - Thousands of leading experts contribute – makes it cutting edge.

Downsides of R

- No Graphical User Interface; though we will use *R Studio* which somewhat simplifies use of *R*.
- Intimidating and picky (everything in *R* is case sensitive!).
- You may have to learn a new language.
- No commercial support or all in one manual.
- Figuring this out can be frustrating (more so in the beginning).

Other R Limitations

- Legacy code/staff:
  - Critical code written in another language & staff familiar with *language X* but not with *R*.
  - Too expensive to rewrite into *R* code.
- Distrust of freeware:
  - "No support! If *SAS* is broken, you call *SAS*. If *R* is broken you call ... ?"
- Efficiency:
  - *R* performs operations on objects in memory. Problems if data size approaches the limits of memory (some work-around's).
- Institutional constraints, e.g.:
  - FDA certification requirements for analysis
  - *http://blog.revolutionanalytics.com/2012/06/fda-r-ok.html*

Resources

- Books:
  - *R for Beginners* by Emmanuel Paradis:
    ftp://cran.r-project.org/pub/R/doc/contrib/Paradis-rdebuts_en.pdf
  - *R and Data Mining*: Examples and Case Studies by Yanchang Zhao:
    ftp://cran.r-project.org/pub/R/doc/contrib/Zhao_R_and_data_mining.pdf
- CRAN (Comprehensive *R* Archive Network):
  - Source of *R* software and documentation: *https://cran.r-project.org/*
- *RStudio:*
  - *R* GUI: *https://www.rstudio.com/products/rstudio/download/*

Help/Manuals

- *An Introduction to R*
  - Very useful, worth reading, only 29 pages long
- *R Reference*
  - Extremely detailed with over three thousand pages!
  - *R* help commands probably more useful.
- *R Data Import/Export*
  - Very useful discussion of an important topic
- *The R Language Definition:* advanced details
- *Writing R Extensions:* advanced (writing packages for CRAN)
- *R Internals:* advanced details
- *R Installation and Administration:* advanced details
- *Sweave User: LaTeX* interface for reproducible research

Dissecting R

- "*R* is a system for statistical computation and graphics"
- Core = interpreted computer language:
  - You type commands and *R* does something
  - Alternative: you can run *R* in batch also
- Built-in support for:
  - Numerous statistical procedures: descriptive stats, linear models, generalized linear models, clustering, etc.
  - Graphics
- External interfaces: files, Internet, databases
- Three main components of *R*:
  - The base packages: always present, define the language,
  - The recommended packages: almost always present, basic analyses
  - Add-on packages: optional, several thousand available at

  http://cran.r-project.org/web/packages/available_packages_by_date.html

What is an R Package?

- An *R* package is a collection of components:
  - Functions,
  - Datasets,
  - Various other things (e.g., package description text objects, object class definitions, etc.)
- Sizes vary enormously:
  - The *beanplot* package has 1 function and 1 description,
  - Some packages have over 1,000 components.