Lecture 1: Introduction

STAT598z: Intro. to computing for statistics

Vinayak Rao

Department of Statistics, Purdue University

Logistics

- **Class** Tue/Thu 1030-1145, Math Building 215
- **Class email** purduestat598z@gmail.com
- **Website** https://varao.github.io/stat598z_fall17/main.html (https://varao.github.io/stat598z_fall17/main.html)
- **Piazza** https://piazza.com/class/ixqh6ogobyf59l (https://piazza.com/class/ixqh6ogobyf59l)
- **Instructor** Vinayak Rao
  - If you email me, include STAT598z in the subject
  - E.g. "STAT598z: My dog ate my homework"
- **Office** Math212
- **Office Hours** 1230 - 1330 Tuesdays or by appointment
- **TA** Boqian Zhang (http://www.stat.purdue.edu/people/graduate_students/)
- **Office Hours** Wednesday 1330-1430

EMERGENCY PREPAREDNESS – A MESSAGE FROM PURDUE

To report an emergency, call 911. To obtain updates regarding an ongoing emergency, sign up for Purdue Alert text messages, view www.purdue.edu/ea.

There are nearly 300 Emergency Telephones outdoors across campus and in parking garages that connect directly to the PUPD. If you feel threatened or need help, push the button and you will be connected immediately.

If we hear a fire alarm during class we will immediately suspend class, evacuate the building, and proceed outdoors. Do not use the elevator.

If we are notified during class of a Shelter in Place requirement for a tornado warning, we will suspend class and shelter in [the basement].

If we are notified during class of a Shelter in Place requirement for a hazardous materials release, or a civil disturbance, including a shooting or other use of weapons, we will suspend class and shelter in the classroom, shutting the door and turning off the lights.

Please review the Emergency Preparedness website for additional information.
http://www.purdue.edu/ehps/emergency_preparedness/index.html

Comp. statistics vs stat. computing

Computational statistics or statistical computing, is that the question?, Lauro, C. , Comp Stat and Data Analysis 23 (1996) (http://econpapers.repec.org/article/eeecsdana/v_3a23_3ay_3a1996_3ai_3a1_3ap_3a191-193.htm)

Statistical Computing: Application of Comp. Sci. to Statistics

- Tools: programming, software, data structures and their manipulation, hardware (GPUs, parallel architectures)
- E.g. Releasing software/ sharing analysis

Computational statistics: Design of algorithms for implementing statistical methods on computers

- Statistical methodology
- E.g. writing a paper for a statistics journal?

This course: more of former

STAT545: more of the latter

Goals of the course

Broadly: to learn programming for Statistics/Data Science

- No programming background required.
- Perhaps not the best class for those already good at this

Our focus will be on

- the R programming language
- *statistical* rather than *general-purpose* computing
- R for reproducible research rather than ad hoc analysis

Topics covered (tentative)

- R fundamentals (data-structures, commands, flow control)
- R packages
- R plotting (ggplot2)
- Debugging with R
- Writing efficient R code
- R Markdown and dynamic documents
- Object-oriented programming
- Functional programming

Advanced topics (depending on how things progress):

- Interactive applications with R shiny
- Introduction to R internals
- Programming with Stan
- Introduction to C -programming, and its R interface

Textbooks

- *"The Art of R Programming: A Tour of Statistical Software. Design"*, Norman Matloff
- *"R for Data Science"*, Garrett Grolemund and Hadley Wickham. (Amazon (http://amzn.to/2aHLAQ1) but also available free (http://r4ds.had.co.nz))

Also useful:

- *"Software for Data Analysis"*, John M. Chambers
- *"An Introduction to R"* (The R manual (http://cran.r-project.org/manuals.html))
- *"Advanced R"*, Hadley Wickham

Grading

- **Homework**: 25%
- **Midterm I**: 25%
- **Midterm II**: 25%
- **Project**: 20%
- **Class participation**: 5%

Homework

(Approximately) weekly assignments

Will involve reading, writing and programming

Are vital to doing well in the exams

Late homework will not be accepted

One (worst) homework will be dropped

You may discuss problems with other sudents, but must:

- write your own solution independently
- name students you had significant discussions with

Purdue's guide on academic integrity (http://www.purdue.edu/odos/osrr/academic-integrity/index.html)

Programming

Central to modern statistics/data analysis. We want:

- computers to do what we don't want to do ourselves
- computers to do what we actually want them to do

Programming involves:

- **Correctness**: getting computers to do what we want
- **Efficiency**: low compute and (more imp.) human time
- **Clarity**: Donald Knuth: "treat a program as … addressed to human beings rather than to a computer"
  - Especially important with messy data

The R programming language

A programming language and environment for statistics

A GNU project available as Free software.

(“Think free as in free speech, not free beer”: Richard Stallman)

You can (and should):

- Install R (available at http://cran.r-project.org/ (http://cran.r-project.org/))
- Look at the R source code
- Modify the R source code (if you’re feeling brave)

You will:

- Write clear, efficient and (hopefully) useful R code

A brief history of R

Based on Bell Labs' S language by John Chambers

Started by Ihaka and Gentleman at the Univ.of Auckland *R: A Language for Data Analysis and Graphics* (http://biostat.mc.vanderbilt.edu/wiki/pub/Main/JeffreyHorner/JCGSR.pdf), (1996)

A high-level interpreted language with convenient features for loading, manipulating and plotting data

A huge collection of user-contributed packages to perform a wide variety of tasks

Widely used in academia, and increasingly popular in industry

The R command prompt

Starting R begins a new session

R presents you with a command prompt or console

Can interact with R through the console:

- Enter command
- R processes command and prints output
- The command q() ends the session

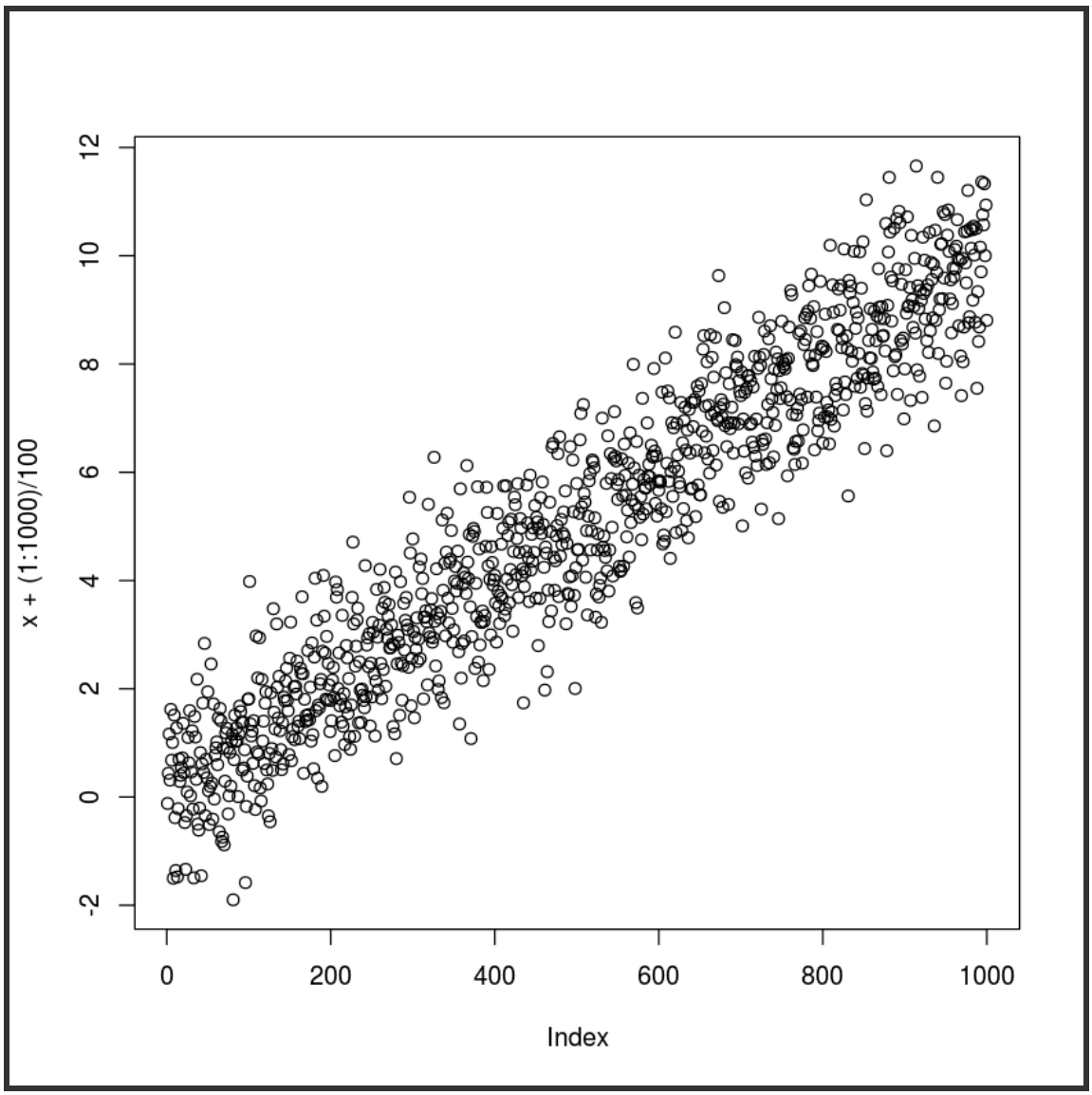
In [1]:
```
1 + 3
```

4

In [4]:
```
x <- rgamma(3,2,1); x # Generate Gamma(2,1) variables
```

3.8421373479265 3.2994556424234 8.04069039166816

In [1]:
```
x <- rnorm(1000)
plot(x+(1:1000)/100)
```



RStudio

RStudio provides a more convenient Integrated Development

Environment (IDE) to interact with R

Layout includes

- an editor
- a console
- workspace/history tabs
- tabs for plots/packages/files etc

Convenient user interface: point-and-click, autocomplete, help etc.

You should install RStudio Desktop (available at rstudio.org)

[RStudio demo]

R Scripting

While we often use R interactively, it is useful to do this through scripts

- Fewer errors
- Better reproducibility
- Can reuse useful sequences of operations
- Can build increasingly complicated sequence of operations

Ultimately, R is a full-fledged programming language for statistical computing: Treat it as such!

R scripting guidelines

Filenames should end with .R (e.g. denoise.R)

Scripts should have explanatory comments

Variables should have informative names

Scripts should be indented appropriately

See R style-guides from:

- Google (https://google.github.io/styleguide/Rguide.xml)
- Hadley Wickham (http://stat405.had.co.nz/r-style.html)

Learning R

We will look up a few useful R packages (e.g. ggplot, plyr)

The next part of the course aims to:

- Write clean, efficient and idiomatic R
- Understand why things done the way they are
- Be comfortable manipulating and presenting data

Dynamic documents and R Markdown

Take the idea of reproducible code to reproducible documents

Instead of working with R commands, work with an entire report

Report includes description of you problem, data and algorithm as well as embedded code and results

You can automatically “compile” the report, which will rerun your code, regenerate your results and form a new report

Allows collaborators to regenerate report on their computer

This is how we will be submitting homeworks

Jupyter notebook

Another nice system for dynamics notebooks is Jupyter notebook

Formerly called ipython notebooks, is still python based, but now supports more languages:

- Ju(lia)Py(thon)R

I made these slides using Jupyter

You can try installing it if you 're prepared to deal with setting up python/python libraries

- Pros: you can play around with these slides
- Cons: I think RMarkdown and knitr is a bit more useful for serious data science

To do

- Install R and RStudio

Reading/Viewing:

- A New York Times article (http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?_r=0)
- Sections 1 to 5 of: A (very) short introduction to R, Torfs and Brauer (http://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf)
- R : the good, the bad and the ugly, John Cook (youtube (https://www.youtube.com/watch?v=6S9r_YbqHy8))
- Nature article on dynamic documents (http://www.nature.com/news/interactive-notebooks-sharing-the-code-1.16261)