STAT 625 Week 1 (a single class)

Jay Emerson and Susan Wang

September 1, 2016

Computing and Reproducible Research

This document was produced in R Studio using R Markdown, choosing the PDF option for the document. You can work without R Studio, but R Studio makes many things like this wonderfully easy. And R Markdown itself is pretty simple (far simpler than full-blown LaTeX). You'll need packages `knitr` and `rmarkdown`. You'll also want LaTeX on your system, though you won't use it directly. In Windows, this is called `MiKTeX`. The full, complete distribution is required (and is quite large, about 2 GB – don't try to download it during class, please). The `MiKTeX` installer will really try to encourage you to install the basic version. It won't work. Trust us, keep your eyes open. On the Mac, you'll need `MacTeX`. It's an easy installation. Anyone using Linux? Probably not, but if so it's likely you can be self-sufficient.

Recommendation: start by installing R (www.r-project.org) and either `MiKTeX` or `MacTeX`. Then install R Studio. We had 110 students last semester, and only 2 problems. These problems arose when something wasn't quite right and (for example), the student tried to recover by modifying the registry in Windows. Don't do crazy stuff like that.

HINT

Work together on computing stuff like this! There are lots of little computing "gotchas" that can be annoying, but fighting through them and succeeding is important. When you graduate and get a real job (or head to graduate school or whatever), being able to solve problems and be computationally self-sufficient will be valuable. Yes, we're here to study statistics (or data analysis, we would prefer to say). But we aren't here just to ***think about*** problems; we're here to ***work on*** problems. Thinking is necessary, but not sufficient.

Markdown?

Looking for information about Markdown? Try http://daringfireball.net/projects/markdown/.

LaTeX?

Let's call this essential if you are doing a PhD – you'll use it for your dissertation and for 95% of the papers you write. But for many it is probably overkill and R Markdown and R scripts will be sufficient. As a side note, you can include many LaTeX things in R Markdown documents.

Toy Example

Let's try a data set that Jay has available on the web. This is a code chunk, which is processed and nicely displayed with the output in the resulting PDF:

```
x <- read.csv("http://www.stat.yale.edu/~jay/diving/Diving2000.csv",
              as.is=TRUE)
dim(x)
```

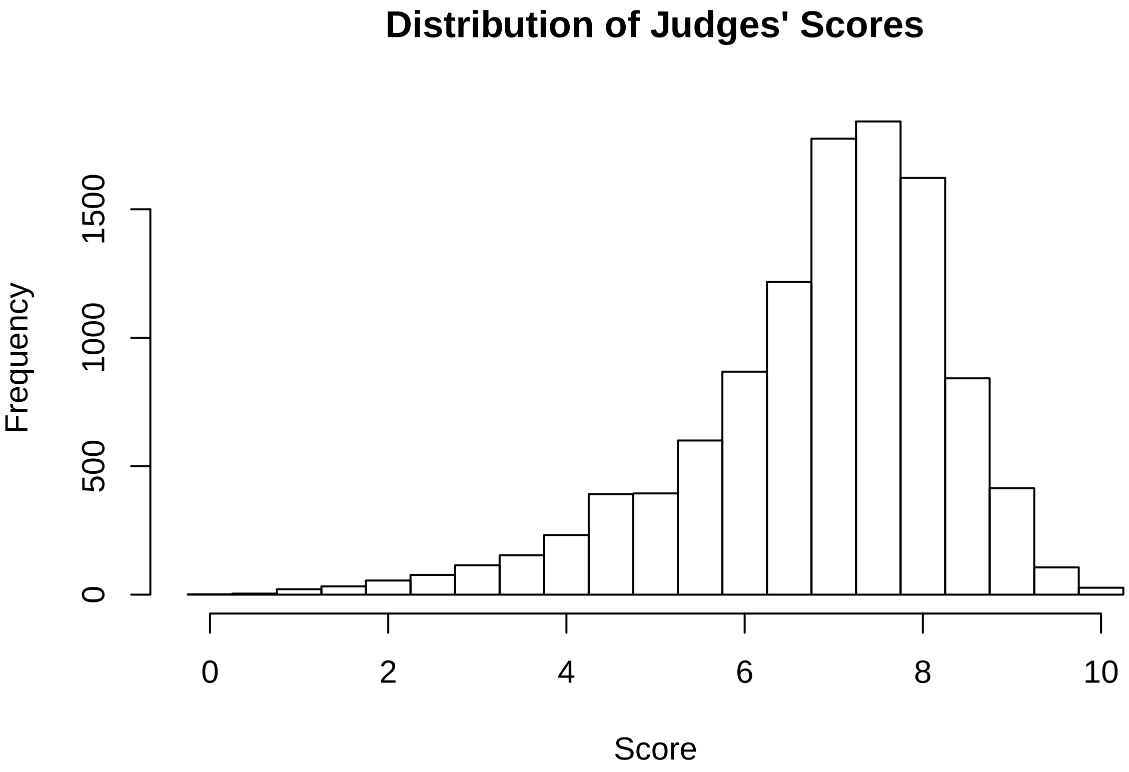
```
## [1] 10787    10
```

```r
mean(x$JScore)
```

```
## [1] 6.832576
```

And R Markdown also makes it easy to integrate graphics. Note that only one graphical display should appear in any one code chunk.

```r
hist(x$JScore, xlab="Score",
     main="Distribution of Judges' Scores",
     breaks=seq(-0.25, 10.25, by=0.5))
```



Reproducible?

This document is entirely ***reproducible*** as long as you have the original R Markdown file, `Y16_W1.Rmd`. You should be able to reproduce this document exactly (or with perhaps superficial formatting differences), including the amazing analysis and plot. Using LaTeX involves a similar workflow that includes code with the document, but is (potentially) more complicated (and hence potentially much more sophisticated). With every passing month I find myself preferring R Markdown.

In this course…

We won't micromanage your computing environment as long as we like the results. We'll pick on you mercilessly for bad code (while trying to maintain a sense of humor about it early in the term). Bad code leads to mistakes and wasted time. We don't like either. And of course we value communication… so effectively, neatly, and concisely presenting your work will be an important component of the course.

TODAY

- Syllabus
- Hoops data scrape example
- Basic New Haven real estate property record toy scrape and challenge
- Homework for Tuesday (yes really)

Homework Due Tuesday 9/6

We will provide the VisionAppraisal files (to be briefly visible and unprotected shortly in Jay's `www.stat.yale.edu/~jay/625/` folder). Please use this version. That is, please don't scrape them directly from the web site – it generates lots of traffic (more than 1.5 GB per batch) and may cause problems.

Create a data frame with three columns (and 27307 rows, because 27307 is the maximum observed parcel id we seem to have this year): parcel id (call it `pid`), raw address (call it `location`), and the 2015 total appraised value (call it `totval`). You'll have a row for each property – if there is no information (or no raw file) there should still be a row for it in the data set (with the `pid` but `NA` values for the other variables). Example: for parcel number 600, the `pid` would be 600, the `location` would be "108 HYDE ST", and the `totval` would be the number 235600 (not the text "$235,600").

Save this data frame as a CSV file "hw1.csv" that **does not have row names**. Upload it to your dropbox on ClassesV2, along with your script that did the work. The comments at the top of the script should acknowledge any collaboration (working together is fine but your script should be your own work and you should understand everything that you do). Intentation and quality of code matters. See the `R_Code_Guidelines.R`, which we're happy to add to.