431 Class 03

Thomas E. Love

2017-09-05

Today's Agenda

1. Project will be discussed on Thursday 2017-09-07
2. Describing data numerically and graphically via R Markdown
    - Brief description of the Getting Started with R examples
    - Live demonstration using the Day 1 survey data
3. Assignment 1 - what should you expect?
4. (maybe) Kidney Cancer death rates

Describing Data - Numerically and Graphically

Numerical quantities focus on expected values, graphical summaries on unexpected values.

-- John Tukey

The Getting Started with R document

R Markdown file and resulting PDF (designed to help you do Assignment 1) describing:

1. A study of chick weights resulting from various feed types
    - Summarizing the distribution of a categorical variable
    - Summarizing the distribution of a continuous variable numerically
    - Graphing the distribution of a continuous variable with a histogram, boxplot and Normal Q-Q plot
2. A study of the growth of orange trees
    - Numerical Summary of two continous variables, and their correlation
    - Scatterplot predicting one continuous variable (outcome) using the other (a predictor)
    - Fitting a Straight-Line model to a scatterplot
    - Stratifying an association on a categorical variable

There are also some tips on getting data into R from Excel.

R Packages

1. **Installing** an R Package is something you need to do once, ideally. (Unless something breaks.)
    - Our Software Installation document has a detailed list of all packages we want you to have installed.
2. **Updating** an R Package is something you do occasionally. (Usually, when something breaks, but I try for once a day.)
    - Update your packages by visiting the Packages tab on the bottom right panel of R Studio and clicking Update.
3. **Loading** an R Package is something you do with the `library` function, in every R Markdown file you ever write.

The key packages for today are `magrittr` and the `tidyverse`. To load them, we'll use:

```
library(magrittr); library(tidyverse)
```

The Class 3 document

Do most of the things that are done in the Getting Started with R document, but

1. build an HTML rather than PDF result
2. build it live in class
3. use our day 1 survey data set instead of a pre-packaged data frame in R.

- You have the complete R Markdown file **Class3_example.Rmd** that I will build in class available to you.

Find `Class3_example.Rmd` on our 431data site, at https://raw.githubusercontent.com/THOMASELOVE/431data/master/Class3_example.Rmd

Class 3 document tasks

1. Download the data called `surveyday1_2017.csv` from our 431 data site
2. Download the `YOURNAME-hw1.Rmd` template we'll use (designed for HW1) also at our 431 data site
3. Open a new project in R using the directory in which you have the data and template.
4. Open the template and edit it a bit to personalize.
5. Load the data using a chunk of R code.
6. Begin exploring the data to address four questions.
    1. How do I summarize a multi-categorical variable, like favorite `color`?
    2. How do I summarize a quantitative variable, like `haircut` price?
    3. What is the relationship between `age` guess and `sex`?
    4. What is the relationship between `pulse` rate and hours of `sleep`?

Discussion of the Class 3 Example

Again, find `Class3_example.Rmd` on our 431data site.

- The raw file is at https://raw.githubusercontent.com/THOMASELOVE/431data/master/Class3_example.Rmd

Assignment 1 (due Friday 2017-09-15)

1. Use the `YOURNAME-hw1.Rmd` template to your advantage.
2. Use the Getting Started in R document from our front page to help guide you.
3. The Course Notes contain all the code you might possibly need.
4. Grading will be very light on this assignment compared to later ones.
5. Submit the assignment (two files: R Markdown, plus either HTML or Word files) via canvas.case.edu
6. Apply the 15-minute rule.
    - If you can't solve a problem in 15 minutes, ask for help.
    - You are **absolutely supposed** to use Google and the TAs (and me) to improve your code.

Kidney Cancer Death Rates

Your map shows U.S. counties.

- The shaded counties are in the top 10% of age-standardized rates for death due to cancer of the kidney/ureter for white males, in 1980-1989.

Your Tasks

1. Describe the patterns you see in the map.
2. Speculate as to the cause of these patterns.

Highest kidney cancer death rates

Lowest kidney cancer death rates

What's next?

- Thursday, we'll be discussing some highlights from Jeff Leek's *The Elements of Data Analytic Style*, chapters 5, 9, 10 and 13.
- **Bring at least one (written down) question and/or comment about something in the text that is meaningful to you.**
    - Chapter 5 is about Exploratory Analysis
    - Chapter 9 is about Written Analyses (keep this in mind for Assignments!)
    - Chapter 10 is about Creating Figures
    - Chapter 13 highlights a few matters of form
    - You may want to read Leek Chapters 1-4 now, too. They're all brief.
- The NHANES data example (Sections 1-6 of the Course Notes)
- Read Silver Introduction and Chapter 1 (about 50 denser pages) by 2017-09-12 (Class 5)
- Assignment 1 is due 2017-09-15 at noon

Notes on the Kidney Cancer example, 1

I first asked you what you noticed about the map, in the hope that someone would point out the obvious pattern, which is that many of the countries in the Great Plains but relatively few near the coasts are shaded.

- Why might that be? Could these be the counties with more old people? Ah, but these rates are age-adjusted.
- They're mostly in rural areas: could the health care there be worse than in major cities? Or perhaps people living in rural areas have less healthy diets, or are exposed to more harmful chemicals? Maybe, but the confusing fact is that the highest 10% and the lowest 10% each show disproportionately higher rates in those Great Plains counties.

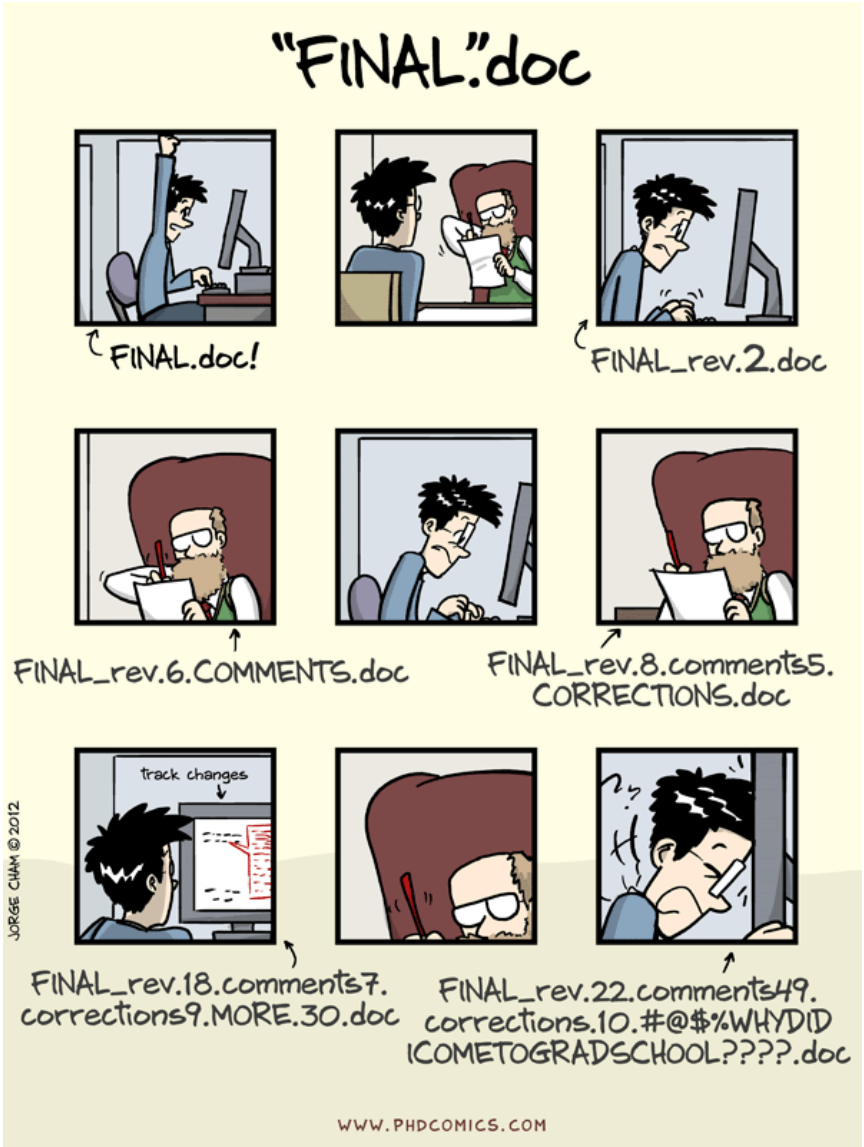
Notes on the Kidney Cancer example, 2

- Consider a county with 100 white males. If it has even one kidney death in the 1980s, its rate is 1 per thousand per year, which is among the highest in the nation. If it has no such deaths, its rate will be 0, which is the lowest in the nation.
- The observed rates for smaller counties are *much* more variable, and hence they are more likely to be shaded, even if nothing special is truly going on.
- If a small county has an observed rate of 1 per thousand per year, it's probably random fluctuation. But if a large county (like Cuyahoga) has a very high rate, it is probably a real phenomenon.

Source

My source for this example was Andrew Gelman and Deborah Nolan's book *Teaching Statistics: a bag of tricks* which is the source of a number of things we'll see in the course, including some of the "age guessing" example we've previously done.

Choose good names for things.



http://www.phdcomics.com/comics/archive.php?comicid=1531

From Karl Broman: Building a Spreadsheet

- Be consistent
- Write dates as YYYY-MM-DD
- Fill in all of the cells
- Put just one thing in a cell
- Make it a rectangle
- Create a data dictionary
- No calculations in the raw data files
- Don't use font color or highlighting as data
- Choose good names for things
- Use data validation to avoid data entry mistakes
- Save the data in plain text files, like .csv

See http://kbroman.org/dataorg/ for more details.