Week 1 Lecture Notes

4 May 2015 – 8 May 2015

1 Principles of Analytic Graphics

Cribbed from "Beautiful Evidence" by Edward Tufte:

1. **Show comparisons.** Always ask, when presented with graphical representations of data, "…compared to what?" Evidence is always relative to another competing hypothesis. For example, don't just show a bar chart of number of symptom-free days of asthma; show a two bar charts comparing symptom-free days between a control group and a group who had air filters in their homes.

2. **Show causality, mechanism, explanation, systematic structure.** Explain how this system is operating. What is the causal framework for thinking about this question? For example, compare the chart of symptom-free days with a chart of level of fine particulate matter (control versus air cleaner).

3. **Show multivariate data.** The world is inherently multivariate. Show as much data in a single chart as possible, to incorporate possible confounders.

4. **Integrate your evidence.** Use words, numbers, images, diagrams. Data graphics should make use of many modes of data presentation; don't let the tool drive the analysis.

5. **Describe and document the evidence.** A data graphic should tell a complete story that is credible. Save the source code, etc.

6. **Content is king.** What's the story we're trying to tell? *Then,* how are we going to present this data?

2 Exploratory Graphs

We use graphs in data analysis to understand data properties, find patterns, suggest modeling strategies, debug an analysis, and communicate results. Exploratory graphs are about the first four of these. Communicative graphs come later.

Exploratory graphs are usually made quickly, in bulk, for personal understanding. How does the data look? Properties, problems? We don't worry about axes/legends. Color or size are primarily used for information.

Example. Air pollution in the United States (EPA National Ambient Air Quality Standards). For fine particle pollution (PM2.5) the “annual mean, averaged over 3 years,” cannot exceed 12 μg/m^3. We want to know: Are there any counties in the U.S. that exceed the national standards for fine particulate pollution?

One-dimensional summaries:

1. Five number summary: Minimum, 1st Q., Median, 3rd Q., Maximum. (`summary`; R also gives the mean.)
2. Box plot (`boxplot`)
3. Histogram (`hist`; can also add `rug`)
4. Bar plot (`barplot`; for categorical variables)

Two-dimensional summaries:

1. Multiple box plots (`boxplot( CAT1   CAT2, ...)`)
2. Scatterplot (`with(DATA, plot(CAT1, CAT2))`, color for more dimensions)
3. Multiple scatterplots

“Quick and dirty,” mostly just use defaults. We can explore basic questions and hypotheses (maybe rule them out).

3 Plotting Systems in R

The Base Plotting System

Uses an “artist’s palette” model: start with a blank canvas and build up from there; first call a plot function, then use annotation functions to add or modify elements (`text`, `lines`, `points`, `axis`).

This is convenient and inutitive, but changes can’t be reverted. It’s also difficult to “translate” to others once the plot has been created (no graphical “language”). The plot is just a series of R commands, so fine control but you have to twiddle everything.

The Lattice System

Plots are created with a single function call (`xyplot`, `bwplot`, etc.). Most useful for conditioning types of plots: looking at how y changes with x across levels of z (sometimes called panel plots). Attributes like margins or spacing are set automatically because the entire plot is specified at once. This system is good for putting many plots on screen at once.

Sometimes it’s awkward to specify an entire plot in a single call. Annotation is not especially intuitive. The use of panel functions and subscripts are difficult to weild and requires much preparation. Can’t change the plot once plotted.

The ggplot2 System

Splits the difference between base and lattice plots. It automatically deals with spacing, text, titles, but also allows you to annotate by “adding” to a plot. It has a superficial similarity to lattice but generally it’s more intuitive to use. It has many defaults, but you can customize as much as you wish.