In-class Data Analysis Module 1:
irregular data, permutation tests and simulation

PUBHLTH 590F: Intro to Stat Computing and Data Visualization (UMass-Amherst)
Instructor: Nicholas G Reich

In-class collaborative exercise

Goals: to practice analysis and visualization techniques on irregular data, and to learn about and implement a permutation test for irregular/mixed data.

You have in your Dropbox folder a dataset, permTestExample.rda. This dataset comes from a setting where measurements were collected from individuals in one of two settings. You could think of it as, for example, a count of organisms observed under two conditions. The "group" variable indicates which group the observation belongs to. The "y" variable represents the count. In more formal statistical terms, you could think of these as samples from two independent and identically distributed samples, where $Y_{1,1}, Y_{1,2}, Y_{1,3}, ..., Y_{1,N_1}$ are the N_1 observations from the first group and $Y_{2,1}, Y_{2,2}, Y_{2,3}, ..., Y_{2,N_2}$ are the N_2 observations from the second group. We can assume that all of the $Y_{1,i}$ observations follow a common, but unknown distribution. Similarly the $Y_{2,i}$ follow a common, but unknown distribution. Our goal is to compare the observed data from the two groups and draw conclusions about how their distributions are similar or different.

Work through the following tasks, documenting your completed work in a reproducible document (using Sweave or knitr).

1. Draw a random sample (without replacement) of 30 observations from this dataset and save it in your workspace.
2. Using just this subsample of the complete dataset, explore the data in R. Generate a few summary statistics and plots and describe the data in a few sentences in non-technical language. What does this data look like? Can you hypothesize about what kind of process may have generated data that look this?
3. **Before running any formal comparison tests**, write down a few methods (at least two) for comparing the two groups that you'd like to run. Justify your choices. Then run those tests.
4. Briefly discuss the results of your analysis.
5. Choose one of the tests you just ran and conduct a permutation test to evaluate the robustness of your results.
6. Briefly discuss the results of the permutation analysis. Do they change your conclusions made above?
7. Summarize your findings in one paragraph. Your target audience: a statistics professor who has not seen this data nor has read the rest of your report.