SDS 325H Exercises 1 · Warm up

Due Monday, January 26

(1) Paint by numbers

(A) Follow along with the `OrchardSprays.R` script, which you can download from the class web site under the "R scripts" tab. This data is from an agricultural experiment that assesses the potency of lime sulphur in repelling honeybees from trees in an orchard. Use the command `?OrchardSprays` to read a detailed description of the data set, which comes pre-installed with R.[1]

Walk through the script to learn some basic commands for plotting and summarizing data. I recommend actually typing the commands in, rather than copying and pasting into the console. Turn in the boxplot of the response (the "decrease" variable) versus lime sulphur concentration. (You can copy and paste R plots directly into a Word, Mac Pages, etc. file.)

[1] "Individual cells of dry comb were filled with measured amounts of lime sulphur emulsion in sucrose solution. Seven different concentrations of lime sulphur . . . were used, as well as a solution containing no lime sulphur. The responses for the different solutions were obtained by releasing 100 bees into the chamber for two hours, and then measuring the decrease in volume of the solutions in the various cells."

(B) Now walk through the `titanic.R` script, which shows you some basic commands for summarizing categorical data. You will also need the `TitanicSurvival.csv` data set. Both the script and data set are available from the class website under the "R scripts" and "Data" tabs, respectively.

The phrase "women and children first" is famously associated with the sinking of the Titanic. Examine the evidence. From what you see in the data on passenger survival, does it seem like women and children survived in disproportionately higher numbers compared to adult men? Include whatever pictures, tables, and summaries you judge to be relevant in making your case. (The script shows you several of these, but you are free to make others.) In particular, if you think the evidence shows that children disproportionately survived the disaster, make sure you consider the boxplot of age versus survival status. You must explain how this plot fits into your understanding of what happened.

(2) Rate of enzymatic reaction versus substrate concentration

The data set `chymotrypsin.csv`[2] contains data on the measured rates of the enzymatic reaction wherein chymotrypsin—an enzyme that is synthesized by the pancreas and found in your digestive system—helps break down protein molecules. For the non-biochemists like me: the rate of any enzymatic reaction obviously depends upon the underlying

[2] On the class website.

substrate concentration. This file contains data from a subset of a larger experiment trying to understand the nature of this relationship for chymotrypsin. This data set has 81 measurements of the reaction rate at 6 different molar concentrations of substrate ("Conc" in the .csv file). The response variable is the measured rate of the reaction. The errors in an experiment like this can be due not merely to imperfect equipment, but also to imperfections in the sample of substrate. (You might be aware that chymotrypsin is highly specific in terms of which amino acids it binds to.)

Examine the relationship between reaction rate versus concentration, treating concentration as a categorical predictor. An important thing to remember here is that concentration is given as a number in the original data set, and so R treat it by default as a numerical variable. If you want to override this behavior and get R to treat a numerical variable as a categorical variable, enclose the name of the variable in the `factor` command. For example, compare the default plots from the following two commands.

```
plot(Rate~Conc, data=chymotrypsin)
boxplot(Rate~factor(Conc), data=chymotrypsin)
```

Describe the overall trend in the data, including appropriate visual and numerical (e.g. groupwise means and standard deviations) summaries. Does the trend look linear or nonlinear? If you feel a bit lost with the R commands here, refer back to the problem on orchard sprays. It has an accompanying R script that will render this problem more or less a matter of plug and play with the new variable names.

(3) Confounding

Read the article entitled "Mom's Meth Use May Affect Kids' Behavior," published on the ABC News website on March 19, 2012[3]. Write a few substantial paragraphs that address the following questions. Do not exceed 1 single-spaced page.[4]

1. What is the primary causal claim made by the researchers who conducted the original study?
2. Are there any issues of confounding and endogeneity that the authors of the original study would have had to consider in making this causal claim? In your estimation, does the author of the popular news article do a satisfactory job of discussing these issues?
3. In perusing the original research article[5], do you get the sense that the study authors addressed these possible confounders?

[3] `http://abcnews.go.com/Health/meth-pregnancy-affect-kids-behavior/story?id=15953718`

[4] This is an upper limit, not necessarily a target!

[5] `http://pediatrics.aappublications.org/content/129/4/681.short`