

**For Tuesday:**

Topic:	Statistical inference
Text:	Scanned copy of chapter 1 in The Statistical Sleuth by Ramsey Schafer

**Exercises:**

1. Environmental Voting of Democrats and Republicans in the U.S. Senate. In this exercise we will consider a data set according to the League of Conservation Voters of pro- and anti-environmental votes cast by U.S. senators during 2005 - 2007. Use visual representation of the data to investigate party differences in the percentage of pro-environment votes.
2. Fish Oil and Blood Pressure. Consider the study from previous week about the correlation between a diet containing fish oil and blood pressure. In the study, the researchers used 7 red and 7 black playing cards to randomly assign the participants to the treatment groups. Does this method constitute a random sampling? Why might the results of this study be important?
3. Write down the names and ages of 10 people. Using coin flips, divide them into two groups, as if for a randomized experiment. Did one group tend to get many of the older subjects? Was there anyway to predict which group would have a higher age in advance of the coin flips?
4. Zinc concentration. An experiment was conducted where a group of rats were split into two groups. One group, group A, received a dietary supplement, while the other group, group B, did not. We are interested in the possible side effect of the treatment. In the data set, the concentration of zinc (in mg/ml) in each rat is listed, as well as the rats' treatment group. Make box-plots for each of the two groups of rats. Preferably in the same plot, and if not then at least with the scale of the axis.

**For Thursday:**

Topic:	(Simple) Linear Regression
Text:	Chapter 2 in Linear Regression -an Introduction to Statistical Models by Peter

**Exercises:**

1. Explain the difference between regression, regression model, and simple linear regression model.
2. Pollen Removal. As part of a study to investigate reproductive strategies in plants, biologists recorded the time spent at sources of pollen and the proportions of pollen removed by bumblebee queens and honeybee workers pollinating a species of lily (Data from L. D. Harder and J. D. Thompson, "Evolutionary Options for Maximizing Pollen Dispersal of Animal-pollinated Plants", *American Naturalist* 133 (1989): 323-44.). Find the data in this weeks folder "data sets".
  - (a) Draw a scatterplot of proportion of pollen removed versus duration of visit, for the bumblebee queens.
  - (b) Find and follow a guide on how to make a simple linear regression line with R.
  - (c) Draw the estimated regression line on the scatterplot.
  - (d) What problems are evident in the residual plot?
  - (e) Try fitting the regression only for those times less than 31 seconds (i.e. excluding the two longest times). Does this fit better?
3. Sampling Bias in Exit Polls. Exit pollsters predict election results before final counts are tallied by sampling voters leaving voting locations. The pollsters have no way of selecting a random sample, so they instruct their interviewers to select every third exiting voter, or fourth, or fifth, or tenth, or some other specific number. Some voters refuse to participate or avoid the interviewer. If the refusers and avoiders have the same voting patterns as the rest of the population, then it shouldn't matter; the sample, although not random, wouldn't be biased. If, however, one candidate's voters are more likely to refuse or avoid interview, the sample would be biased and could lead to misleading conclusion.

On November 4, 2004, exit pollsters incorrectly predicted that John Kerry would win the U.S. presidential election over George W. Bush. The exit polls overstated the Kerry advantage in 42 of 50 states. No one expects exit polls to be exact, but chances alone cannot reasonably explain this discrepancy. Although fraud is a possibility, the data are also consistent, with Bush supporters being more likely than Kerry supporters to refuse or avoid participation in the exit poll.

In a postelection evaluation, the exit polling agency investigated voter avoidance of interviewers. The data set includes the average Kerry exit poll overestimate (determined after the actual counts were available for a large number of voting precincts, grouped according to the distance of the interviewer from the door. If Bush voters were more likely to avoid interviewers in general, one might also expect a greater avoidance with increasing distance to the interviewer (since there is more opportunity to escape). A positive relationship between distance of the interviewer from the door and amount of Kerry overestimate, therefore, would lend credibility to the theory that Bush voters were more likely to avoid exit poll interviews. How strong is the evidence that the mean Kerry overestimate increases with increasing distance of interviewer from the door? (Data from Evaluation of Edison/Mitofsky Election System 2004 prepared by Edison Media Research and Mitofsky International for the National Election Pool (NEP), January 15, 2005)

4. Sampling Bias in Exit Polls II. This exercise is about differential interview refusal rates in the exit polls conducted in the 2004 U.S. presidential election. The data shows the average proportion of voters who refuse to be interviewed at precincts grouped according to the approximate age of the interviewer. What evidence do the data provide that the mean refusal rate decreased with increasing age of interviewer?