Data Sets

This class will be more interesting and useful for you if you analyze data that you are interested in. Also, many of the ideas we'll discuss hinge on subtleties that can be hard to understand if you haven't collected data before. Mathematicians and statistics are at a disadvantage in that we rarely get the chance to collect our own data. For this reason, you'll do a short project in which you'll provide three data sets for analysis in this class.

1. Controlled Study
Collect data involving at least two variables in which at least one of the variables is "controlled." This means you determine the x value and record the resulting y-value. You should record at least 3 y-values for each setting of the x-value. For example, you might measure the time it takes a marble to roll down a chute (y) when the chute is set at various angles (x).

2. Observational Study
Collect data involving at least two variables in which you did not control the x variable(s). For example, the number of pages in a set of books and the width, length, and depth of each book.

3. A data set from a real study.
This might be from your own research or from someone else's. It can be (and probably will have to be) a previously published data set, and can even be previously analyzed. Think of something you're interested in, and try to find data on that topic. One good place to start:

http://it.stlawu.edu/~rlock/tise98/onepage.html#Section%207

<u>Rules.</u>

At least one data set must be multivariate.

Don't worry about sample size. Collect enough to make it interesting and useful to you.

Each data set should consist of two files. The first contains the data in text-format, tab-delimited, with variable names on the first row. Each row represents an observation; each column represents a variable. Missing values should be left blank. The second file will explain and describe the data set. This file must answer these questions:

> Who? What? Where? When? Why? and How? What research question could the data be used to answer? Who or what was measured? What are the units? How were the measurements collected? When and where were they collected? What do the variable names mean?

No time series. This means the 'x variable' can't be the date or elapsed time.

Both files must be submitted via email in electronic form (text only format). Files will be posted on the class web page so that they can be shared.