36-315: Statistical Graphics and Visualization

Lab 1
Date: January 14, 2003 **Due: end of lab**

(This lab was written for R but today we will use S-PLUS, which is nearly the same. The only changes are to steps 4 and 5.)
The data is currently an Excel spreadsheet and needs to be converted to a format R can read.

1. Download `lab1.s` and `lab1.xls` from the course web page:

   `http://www.stat.cmu.edu/~minka/courses/36-315/lab/`

   Save in the `My Documents` folder.

2. Open `My Documents` and double-click on `lab1.xls` to start Excel.

3. Use `File -> Save As...` to save the data in Comma-Separated Value (csv) format. In response to the dialogs, click `Ok` then `Yes`.

Start S-PLUS

4. `Start -> Programs -> Math & Stats -> S-PLUS 6.1`

   A small window will appear asking for a folder name. Browse to `My Documents`.

5. Load the special functions for this lab by typing in the Command Window:

   ```
   source("lab1.s")
   ```

   and press enter. If this doesn't work, the file may have been renamed to `lab1.s.txt` when you downloaded it.

6. Arrange the various sub-windows by selecting the menu option `Window -> Tile Vertical`. If you wish, you can delete the Object Explorer sub-window.

Load the data

7. Read the csv file into an R data frame, by typing into the command window:

   ```
   frame = read.csv("lab1.csv")
   ```

   (If you mistype a command, you can always press up arrow and reenter it.)

8. To check that the data was loaded correctly:

   ```
   summary(frame)
   ```

9. Extract the column `POPPSQMI` from `frame`:

   ```
   x = frame[,"POPPSQMI"]
   ```

The data is now a vector of numbers called `x`. It is the population per square mile (population density) in each census tract in Pennsylvania, according to the 1990 census.

Make plots

10. Using the handout on the `hist` function, make a histogram of `x`, using the default number of bins, with the axis label "Population per square mile".

11. Now change the histogram to have 1000 bins and zoom in on the range 0...3000. Notice that the density falls quickly in the beginning and then levels off, producing a "knee" in the histogram, where the slope abruptly changes.

12. Create a new plot window by typing `windows()`. Now make a histogram of `log(x)` instead of `x`, using the default number of bins. (Type `log(x)` wherever you would normally type `x`. S-PLUS can evaluate any mathematical expression.)

13. Create a new plot window and make a histogram of `log(x)` using 70 bins. Notice the difference between the previous graph.

14. Show us the last three graphs and answer the following question: where did the "knee" in the first plot move to in the third plot?