# 36-315: Statistical Graphics and Visualization

**Homework 4**
**Date: February 3, 2002** **Due: start of class February 10, 2002**

---

1. In this problem, you will write some R code, which you should submit along with your plots. Download `hw4.rda` and source `lab4.r` just like in lab. Loading the data into R will define a table `age.vs.density`, which is the number of people in different age groups across Pennsylvania, split according to population density. It also defines a table `width` which stores the width of each age interval.

   (a) Compute a table containing the probability of each age interval in each population density group. Then divide each element of this table by the corresponding element of the `width` table, to standardize the probability of each age interval (otherwise large intervals will have an undue share of probability). The "/" symbol can divide entire tables, just as it can divide numbers. Assign the standardized table to a variable. (Hint: the whole thing is two lines of code.)

   (b) Using the standardized table, make a pie chart comparison of the age distributions in high, medium, and low population density (city, suburbs, and country).

   (c) Make two bar charts with side-by-side bars: one with bars grouped by age and the other grouped by population density.

   (d) Make a line chart where the horizontal axis is age. If there is a dip at ages 10–13, then you didn't standardize the intervals properly.

   (e) Which single plot gives the best overall picture of the differences between the three population groups? Feel free to try other plots than the four above.

   (f) The age intervals in the data are somewhat arbitrary. Describe an alternative division of ages into just five intervals, which would preserve the main differences between groups that you see in the plots.

   (g) In terms of your new age intervals, summarize the difference in age distribution between high, medium, and low population density.

   (h) One way to explain the differences is mobility—people of a certain age tend to move into certain areas. Assuming this is true, describe what the plot says about how a person moves around as they go through life.

2. A recent New York Times article, reprinted on the next page, used an interesting variety of chart. You might call it a "square chart." It is a type different than any we've discussed in class, however you can still apply the principles of encoding, visual connection, and data-ink to analyze it.

   (a) Which one of the six visual encodings discussed in class (angle, shading, texture, area, length, position) does this chart use? (Ignoring the text on the side of the chart.)

   (b) Sketch a modification of the chart which uses the same style and encoding but less ink.

   (c) According to principles of visual encoding, visual connection, and data-ink, how does the published chart rank among those discussed in class? Is it among the best or among the worst?

3. The "fourfold display", illustrated on page 4, has been proposed as a visual test for independence in a $2 \times 2$ table of counts (Friendly, 2000). Under independence, the quadrants would be all the same size so the shape would be a circle. Because it is not a circle and the differences exceed the confidence intervals, we can conclude that the variables (sex and admission) are not independent.

   (a) Critically evaluate this display in terms of its choice of visual encoding, visual connections, and use of ink.

   (b) Testing for independence in this table really boils down to comparing two numbers: the probability of admission for males and the probability of admission for females. Independence requires that they are the same. Sketch an alternative display for $2 \times 2$ tables which how these two quantities differ, and whether the difference is statistically significant. Your display should try to optimize the principles of encoding, connection, and ink.