

Multivariate Data Visualization

Summary: We consider how to use Scheme and GIMP to visualize statistical data.

Contents:

- Introduction: Data Visualization
- Background: Characterizing Data
- Visualization Characteristics

Introduction: Data Visualization

We have explored GIMP primarily as a platform for making visually interesting images (calling them art would probably be a stretch). However, it is also possible to use the GIMP (or any graphics platform) to make *informative* images. In particular, many statisticians have found that complex data sets are easier to understand when the data are represented visually in addition to textually.

For example, in a data set with one independent variable (say, the size of an input for an algorithm) and one dependent variable (say, the number of calls to `CONS` that the algorithm makes on one input of that size), we might plot a graph to better understand that relationship. If the points form a straight, but not horizontal, line, we might conclude that the relationship is linear. If the points form a more complex shape, we need to do more analysis.

One particularly challenging aspect of data visualization is what to do with *multivariate* data, data in which we have more than two variables, and in which it is not necessarily clear which variables are dependent and which variables are independent. In such situations, one needs to explore how best to represent the various variables for each observation.

For example, consider the wealth of information about countries of the world available at <http://devdata.worldbank.org/data-query/>. There are a few hundred countries represented, fifty four possible attributes (amount of agricultural land, population, life expectancy, per-capita income, etc.), and six years of data (2000 to 2005, inclusive).

For these data, someone might be interested in exploring relationships that involve population, life expectancy, per-capita income, and year. How can we represent these data to best explore these relationships?

Background: Characterizing Data

Before we delve into the particular techniques one might use to visualize data, let us consider first how statisticians tend to characterize data. The broadest characterization that statisticians make is between categorical data and numerical data.

Categorical data are data that have been broken up into categories, and in which the categories have no numerical relationship, other than set-theoretic relationship. For example, we would call gender (male or female) a categorical variable and (except for some outliers that we'll ignore for now), the relationship between the two is that anyone who is not a woman is a man, and vice versa. For the World Bank data described in the introduction, one obvious categorical variable is the continent to which the country belongs.

Quantitative data are data that are represented numerically, and that often therefore have a numeric relationship. For example, for numeric grades, we might say that a grade of 100 is twice as large as a grade of 50. For the World Bank data described in the introduction, most variables are quantitative.

While it is not generally possible to represent categorical data quantitatively (except, of course, in terms of the number of data points that fall in each category), it is straightforward to represent quantitative data categorically. For example, we might separate grades into those below 60 (failing) and those above 60 (passing). Sometimes, we will categorize less broadly (e.g., A, B, C, D, and F).

At times, we may enumerate categories (e.g., Male is category 0, Female is category 1, No Answer is category 2, and Other is category 3). However, such enumeration should never be taken to mean that there is any real quantitative aspect to the data. For example, we would never average genders (nor continents nor ...).

Visualization Characteristics

In general, we think about representing each data point with a single symbol. We map different attributes of the point to different attributes of the symbol.

For categorical attributes, we can associate a color, shape, or pattern with each category. For a student, we might use one shape (e.g., a circle) for humanities majors, another (e.g., a square) for science majors, and a third (e.g., a triangle) for social studies majors. We might then use color to represent the dormitory in which the student resides (e.g., red for North Campus, yellow for South Campus, blue for East Campus, and purple for off campus). Experienced designers pay attention to those who are color blind or color deficient in choosing colors, but we leave that topic for another course.

For quantitative attributes, we might map the number to the size of the symbol, the shade of the color (lower numbers lighter, higher numbers darker), the weight of the border of the shape, the x position of the center, or the y position of the center.

In the corresponding laboratory, you will have a chance to explore some of these variations.

Copyright © 2006 Samuel A. Rebelsky. This work is licensed under a Creative Commons Attribution-NonCommercial 2.5 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/2.5/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.