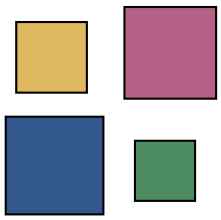


Prof. Antony Unwin
Dept of Computer-Oriented Statistics and Data Analysis
Institute for Mathematics
University of Augsburg
http://stats.math.uni-augsburg.de/

Stochastik IV – Graphical Data Analysis

Exercise Sheet 2: Graphics for continuous variables

Tutorial: Monday, 31st October, 2011, 12.15 - 13.45 Uhr, Room 2001 T

1. **Galaxies**
   The dataset is called *galaxies* in the package *MASS*.
   (a) Draw a histogram, a boxplot, and a density estimate of the data. What information can you get from each plot?
   (b) Experiment with different binwidths for the histogram and different bandwidths for the density estimates. What choices do you think are best for conveying the information in the data?
   (c) How many plots do you think you need to present the information? Which ones(s)?

2. **Student Survey**
   The dataset is called *survey* in the package *MASS*.
   (a) Draw a histogram of student heights and overlay a density estimate of the data. Is there evidence of bimodality?
   (b) Experiment with different binwidths for the histogram and different bandwidths for the density estimates. What choices do you think are best for conveying the information in the data?
   (c) Compare male and female heights using separate aligned density estimates that are common scaled.

3. **Zuni Educational Funding**
   The *zuni* dataset in the package *lawstat* seems quite simple. There are three pieces of information about each of 89 school districts in the US State of New Mexico: the name of the district, the average revenue per pupil in dollars and the number of pupils. This apparent simplicity hides an interesting story. The data were used to determine how to allocate substantial amounts of money and there were intense legal disagreements about how the law should be interpreted and how the data should be used. Gastwirth was heavily involved and has written informatively about the case from a statistical point of view. One statistical issue was the rule that in determining whether district revenues were sufficiently equalised, the 5th and 95th percentiles of the distribution should be compared.
   (a) Are the lowest and highest 5% of the revenue values extreme? Do you prefer a histogram or a boxplot for showing this?
   (b) Having removed the lowest and highest 5% of the cases, draw a density estimate of the remaining data and discuss whether the resulting distribution looks symmetric.
   (c) Draw a QQ-plot for the data after removal of the 5% at each end and comment on whether you would regard the distribution as normal or not.

4. **Non-detectable**
   The dataset *CHAIN* from the package *mi* includes data from a study of 532 HIV patients in New York. The variable `h39b.W1` records, according to the R help page, 'log of self reported viral load level at round 6th (0 represents undetectable level)'. Using the function `table(CHAIN$h39b.W1)` you can find out that there are 188 cases with value 0, i.e. with undetectable levels. Further examination of the dataset (using `mi.info(CHAIN)`, for instance) reveals that additionally 179 cases have missing values for `h39b.W1`. What plots would you draw to show the distribution of the variable `CHAIN$h39b.W1` (i) with the 0 values (ii) without the 0 values?

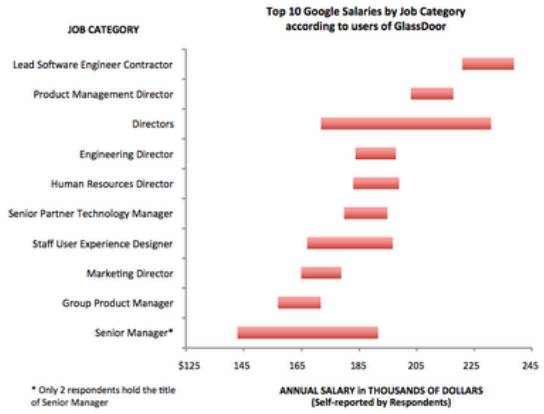
Extra: A Graphic from the Media about Google Salaries

The massive burden of pie charts

The world would be a better place if pie charts were banned. I respect the creativity that went to producing charts like the one below but surely, pie charts don't deserve that level of attention:

(Via Business Insider via Jobvine)

Here is the same information using a bar chart.



It's amazing how much extra stuff (colors, text, shapes, lines, labels, etc.) is needed when you use pie charts.

Figure 1: The website Junk Charts comments on a chart of Google Salaries prepared by Jobvine (Source: junkcharts.typepad.com)

Do you think Junk Charts criticism is sound? Is the second graphic the best that could be done or does it have weaknesses too? (You will probably prefer to look at the displays on the website.)