Letter-Value Box Plots: Adjusting Box Plots for Large Data Sets

Heike Hofmann, Karen Kafadar, Hadley Wickham

January 26, 2006

Abstract

Conventional boxplots (Tukey 1977) are useful displays for conveying rough information about the central 50% of the data and the extent of the data. For moderate-sized data sets ($n < 1000$), detailed estimates of tail behavior beyond the quartiles may not be trustworthy, so the information provided by boxplots is appropriately somewhat vague beyond the quartiles, and the expected number of "outliers" and "far-out" values for a Gaussian sample of size n is often less than 10 (Hoaglin, Iglewicz, and Tukey 1986). Large data sets ($n \approx 10,000 - 100,000$) afford more precise estimates of quantiles in the tails beyond the quartiles and also can be expected to present a large number of "outliers" (about $0.4 + 0.007n$). The letter-value box plot addresses both these shortcomings: it conveys more detailed information in the tails using letter values, only out to the depths where the letter values are reliable estimates of their corresponding quantiles (corresponding to tail areas of roughly 2^{-i}); "outliers" are defined as a function of the most extreme letter value shown. All aspects shown on the letter-value boxplot are actual observations, thus remaining faithful to the principles that governed Tukey's original boxplot. We illustrate the letter-value boxplot with some actual examples that demonstrate their usefulness, particularly for large data sets.

Key words: boxplots, quantiles, letter value display, fourth, order statistic, tail area