Journal of Statistical Software

November 2008, Volume 28, Code Snippet 1. *http://www.jstatsoft.org/*

Beanplot: A Boxplot Alternative for Visual Comparison of Distributions

Peter Kampstra
VU University Amsterdam

Abstract

Boxplots and variants thereof are frequently used to compare univariate data. Boxplots have the disadvantage that they are not easy to explain to non-mathematicians, and that some information is not visible. A beanplot is an alternative to the boxplot for visual comparison of univariate data between groups. In a beanplot, the individual observations are shown as small lines in a one-dimensional scatter plot. Next to that, the estimated density of the distributions is visible and the average is shown. It is easy to compare different groups of data in a beanplot and to see if a group contains enough observations to make the group interesting from a statistical point of view. Anomalies in the data, such as bimodal distributions and duplicate measurements, are easily spotted in a beanplot. For groups with two subgroups (e.g., male and female), there is a special asymmetric beanplot. For easy usage, an implementation was made in R.

Keywords: exploratory data analysis, descriptive statistics, box plot, boxplot, violin plot, density plot, comparing univariate data, visualization, beanplot, R, graphical methods, visualization.

1. Introduction

There are many known plots that are used to show distributions of univariate data. There are histograms, stem-and-leaf-plots, boxplots, density traces, and many more. Most of these plots are not handy when comparing multiple batches of univariate data. For example, comparing multiple histograms or stem-and-leaf plots is difficult because of the space they take. Multiple density traces are difficult to compare when there are many of them plotted in one plot, because the space becomes cluttered. Therefore, when comparing distributions between batches, Tukey's boxplot is commonly used.

There are many variations of the boxplot. For example, a variable-width notched box-plot

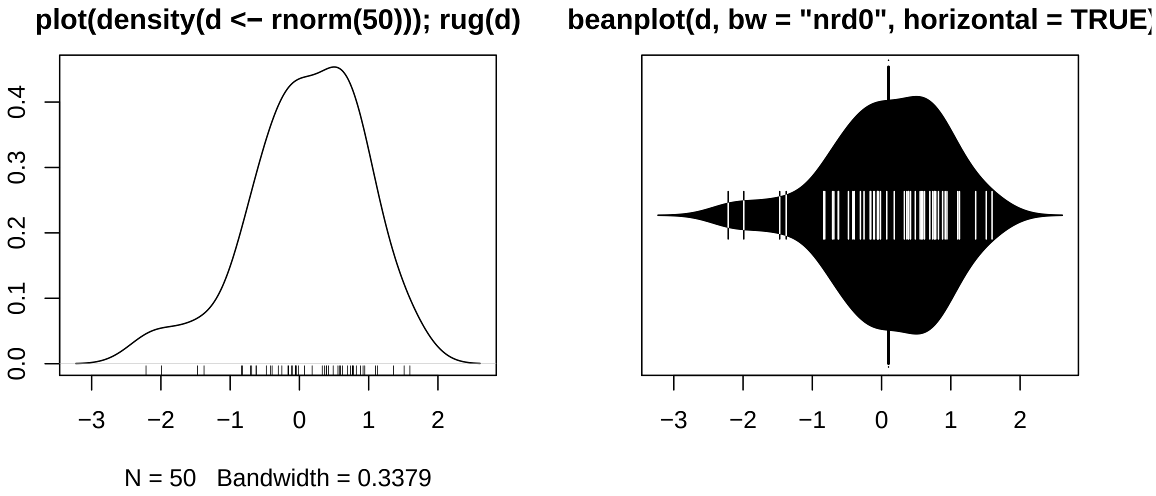

Figure 1: A density trace of a normal distribution with a `rug` (1d-scatter plot) and its corresponding beanplot. The small lines represent individual data points.

(McGill, Tukey, and Larsen 1978) shows the number of observations in a batch using the width of the box, while the notches give an indication of the statistical difference between two batches. Another variation of the boxplot is the violin plot described in Hintze and Nelson (1998), in which a density trace is combined with the quartiles of a boxplot. Individual outliers are not visible in a violin plot.

For smaller datasets there is an alternative to the boxplot, namely a one-dimensional (1d) scatter plot, or stripchart. In such a plot, one dot is plotted for each observation. This alternative to the boxplot is sometimes used (e.g., Box, Hunter, and Hunter 1978), but it only works if there are very few points per batch, because no summarization is provided.

All of the known methods for comparing distributions between batches suffer from some problems. A 1d-scatter plot is only useful in case there are very few values per batch. A boxplot uses quartiles, which are difficult to explain to non-mathematicians (e.g., Bakker, Biehler, and Konold 2005). Next to that, the detection of outliers is quite arbitrary, especially in case of non-normal underlying distributions. Even for normal distributions the number of outliers detected will grow if the number of observations grows, which makes individual outliers undetectable. In a violin plot the underlying distribution is more visible, but individual data points besides the minimum and maximum are not visible at all and no indication of the number of observations in a group is given.

This article therefore proposes a combination between a 1d-scatter plot and a density trace, which is called a beanplot. In such a plot, outliers do not have to be detected, because all individual observations are visible in the scatter plot. Slightly complicated concepts such as quartiles are not used, but instead simply the average is used to summarize the batches. Next to that, a density trace similar to the violin plot is used to summarize the distribution of the batches. The rest of this article explains the details of the beanplot and the implementation in R (R Development Core Team 2008). The implementation is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=beanplot`. Next to that, some examples are shown, like comparisons with a boxplot and a violin plot.

2. The beanplot

A beanplot is a plot in which (one or) multiple batches (”beans”) are shown. An example of such a bean is shown in Figure 1. Each bean consists of a density trace, which is mirrored to form a polygon shape. Next to that, a one-dimensional scatter plot shows all the individual measurements, like in a stripchart. The scatter plot is drawn using one small line for each observation in a batch. If a small line is drawn outside of the density shape, a different color is used to draw the line. This ensures that the density of a batch is still visible, even if there are many small lines that fall partly outside the density shape. To enable easy comparison, a per-batch average and an overall average is drawn (see Figure 3 for an overall line). For groups with two subgroups (e.g., male and female), there is a special asymmetric beanplot.

The name The name beanplot stems from green beans. The density shape can be seen as the pod of a green bean, while the scatter plot shows the seeds inside the pod.

The density shape The density shape used is a polygon given by a normal density trace and its mirrored version. Such a polygon often looks a bit like a violin, and is also used in a violin plot. In R a density trace can be computed by using `density`. For computing such a density trace, a bandwidth has to be selected. Per default, the implementation of `beanplot` uses the Sheather-Jones method to select a bandwidth per batch, which seems to be preferred and close to optimal (Venables and Ripley 2002, page 129). The bandwidths per batch are averaged over all batches in order to have a fair comparison between batches.

The use of the same bandwidth for all beans has a small side-effect for batches that contain few data points. In such a case, the width of such a bean can become quite huge, drawing attention to a less-interesting bean in terms of statistical significance. To overcome this problem, the width of beans with fewer than 10 data points is scaled linearly (so a bean with 3 data points is only 3/10 of its normal width).

1d-scatter plot and multiple equal observations Combining a density trace with a 1d-scatter plot is frequently done. The left side of Figure 1 shows an example produced by R. In this figure the corresponding beanplot is also shown, which is a simple manipulation of the density plot. The most complicated thing that needs to be done is the change in color when a scatter plot line becomes outside of the polygon shape. Fortunately, the crossings can simply be calculated by linear interpolation (`approx` in R). In case there are multiple observations with the same value in a batch, the individual small lines are added together, increasing the length of the line (see Figure 3 for an example). Therefore, duplicate observations are easily spotted. Observations that are almost equal to each other can also be spotted if transparent drawing is used (in R this only works on some devices, currently `pdf` and `quartz`).

The (overall) average While a boxplot and its variants make use of the median of a group of data points, per default a beanplot uses the average of the group and also shows an overall average. This is because an average is easier to explain to non-mathematicians, and an average usually gives useful information if a density trace is useful.

Asymmetric beanplots Normally, the beans are symmetric to compare them easily. Sometimes, the group data being analysed contains two subgroups for each group, for example male and female subjects. In these situations, each subgroup can take one side of a complete bean (see Figure 4 for an example).

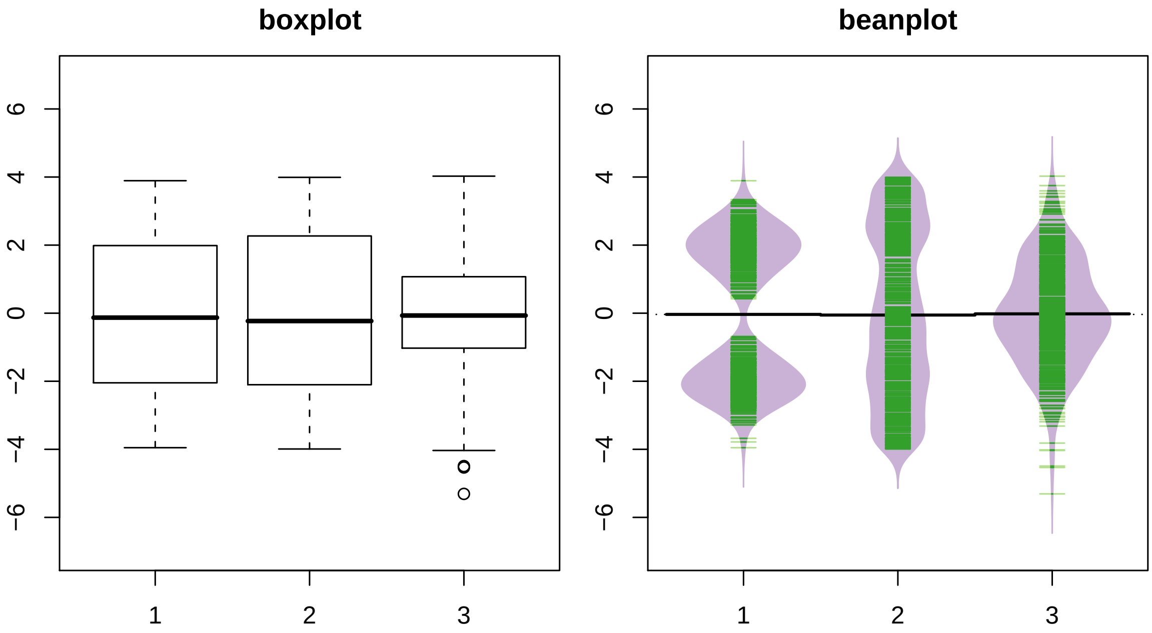

Figure 2: Plots for a bimodal, a uniform and a normal distribution. In the beanplot the green lines show individual observations, while the purple area shows the distribution.

3. Examples of usage

3.1. Some distributions

In Figure 2 some distributions are drawn in a boxplot and in a beanplot. In the boxplot the location of the quartiles does not clearly indicate a difference between the distributions, while the beanplot clearly shows the difference between the distributions. The colors are compatible with the paired colorset from **RColorBrewer** (Neuwirth 2007). The plot was produced using the following code:

```
R> library("beanplot")
R> par(mfrow = c(1, 2), mai = c(0.5, 0.5, 0.5, 0.1))
R> mu <- 2
R> si <- 0.6
R> c <- 500
R> bimodal <- c(rnorm(c/2, -mu, si), rnorm(c/2, mu, si))
R> uniform <- runif(c, -4, 4)
R> normal <- rnorm(c, 0, 1.5)
R> ylim <- c(-7, 7)
R> boxplot(bimodal, uniform, normal, ylim = ylim, main = "boxplot",
+    names = 1:3)
R> beanplot(bimodal, uniform, normal, ylim = ylim, main = "beanplot",
+    col = c("#CAB2D6", "#33A02C", "#B2DF8A"), border = "#CAB2D6")
```

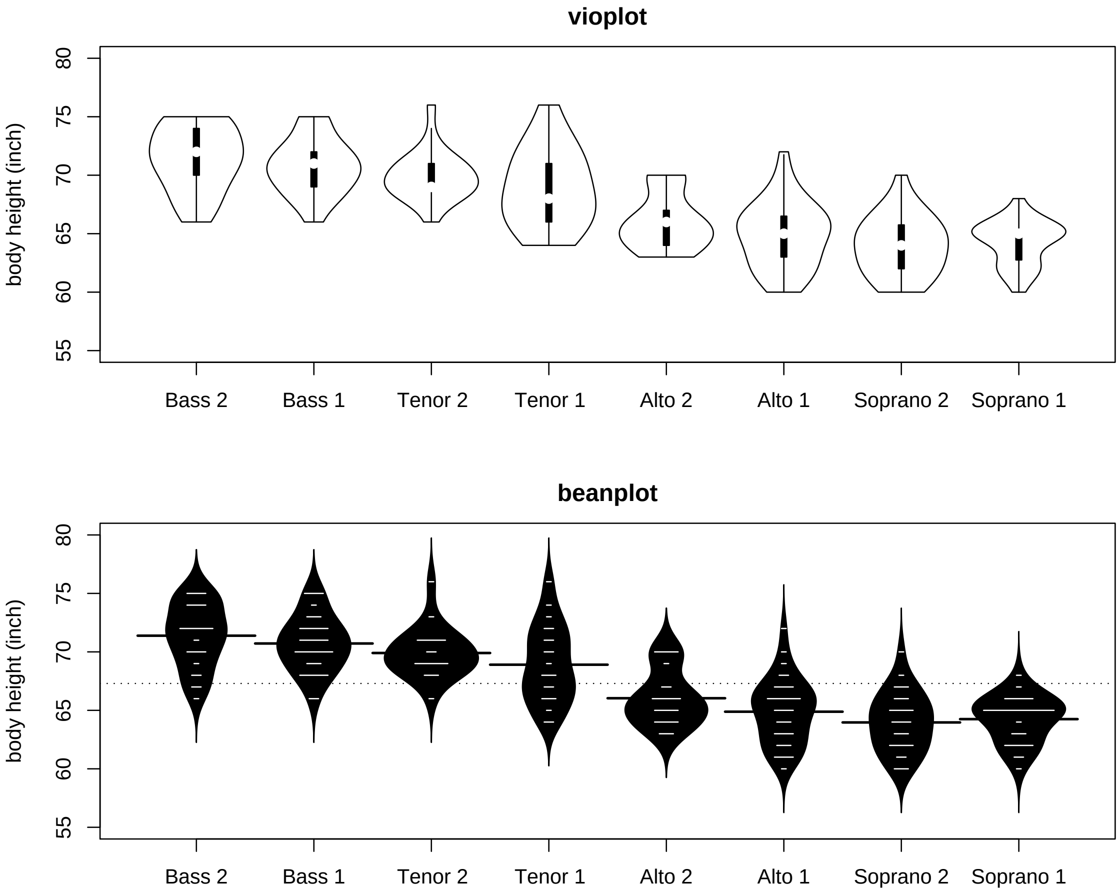

Figure 3: The height in inches of different singers. The beanplot clearly shows the individual measurements were rounded to whole inches.

3.2. Singers

Figure 3 shows a violin plot (as produced by **vioplot**, Adler 2005) and a beanplot for the body heights of different singers taken from Chambers, Cleveland, Kleiner, and Tukey (1983) and available in the **lattice** package (Sarkar 2008). While the violin plot also clearly shows that different groups of singers appear to have different body heights, the beanplot shows extra information. In the beanplot it is visible that the measurements are in whole inches, and that there were many singers with a height of 65 inches in group Soprano 1. Also, an indication of the number of measurements is visible. The plots were produced by using the following code:

```
R> library("vioplot")
R> data("singer", package = "lattice")
R> ylim <- c(55, 80)
R> par(mfrow = c(2, 1), mai = c(0.8, 0.8, 0.5, 0.5))
R> data <- split(singer$height, singer$voice.part)
R> names(data)[1] <- "x"
```

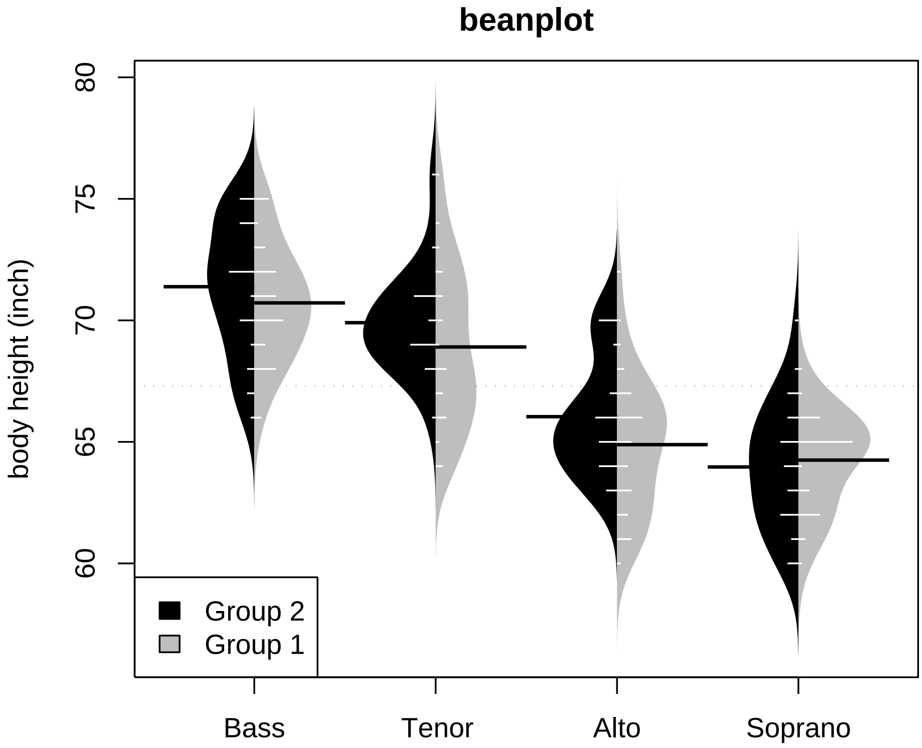

Figure 4: An asymmetric beanplot of the singers.

```
R> do.call("vioplot", c(data,
+    list(ylim = ylim, names = levels(singer$voice.part), col = "white")))
R> title(main = "vioplot", ylab = "body height (inch)")
R> beanplot(height ~ voice.part, data = singer, ll = 0.04, main = "beanplot",
+    ylim = ylim, ylab = "body height (inch)")
```

The dataset contains two subgroups per group, and is therefore suitable for an asymmetric version, which is shown in Figure 4. This plot was produced by the following code:

```
R> par(lend = 1, mai = c(0.8, 0.8, 0.5, 0.5))
R> beanplot(height ~ voice.part, data = singer, ll = 0.04,
+    main = "beanplot", ylab = "body height (inch)", side = "both",
+    border = NA, col = list("black", c("grey", "white")))
R> legend("bottomleft", fill = c("black", "grey"),
+    legend = c("Group 2", "Group 1"))
```

3.3. Easy usage in R

The implementation of `beanplot` in package **beanplot** has kept easy usage in mind. It is compatible with similar functions like `boxplot`, `stripchart`, and `vioplot` in package **vioplot**

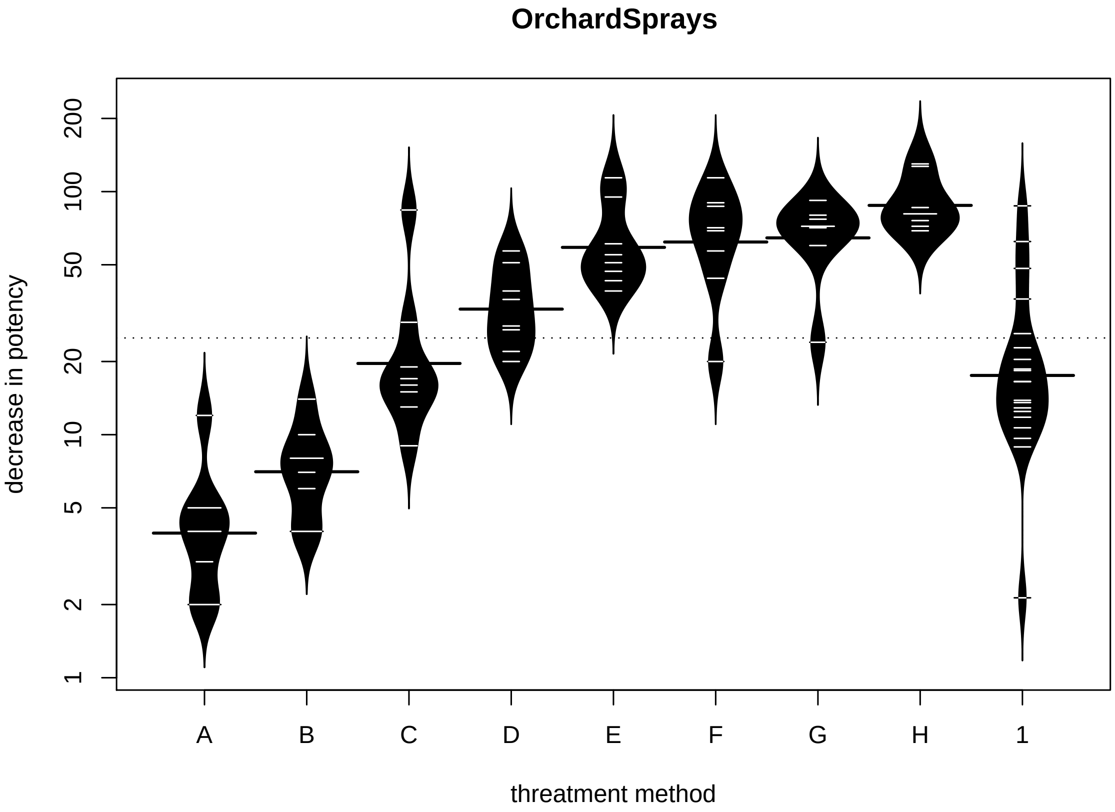

Figure 5: Comparing the potency of various constituents of orchard sprays in repelling honeybees for different threatments with a normal distribution (method '1').

(Adler 2005). Next to that, the **beanplot** package also supports usages that are not possible with these commands. For example, it is possible to combine formulas and vectors as input data if an user wants to compare some things quickly. Therefore, the following code works if the user wants to visually compare data from a formula with a generated normal distribution:

```
R> beanplot(decrease ~ treatment, data = OrchardSprays, exp(rnorm(20, 3)),
+    xlab = "threatment method", ylab = "decrease in potency",
+    main = "OrchardSprays")
```

The results are shown in Figure 5. As an additional aid to the user, a log-axis is automatically selected in this case by checking the outcomes of a `shapiro.test` and the user is notified about this. In case of a log-axis, the density trace is computed using a log-transformation and the geometric average is used instead of the normal average. Therefore, using `beanplot` with a lognormal distribution on a log-axis does not produce strange results, like the direct usage of `boxplot` does, which will show lots of 'outliers' in this scenario.

4. Conclusions

This article showed that a beanplot is a plot that is easy to explain, and enables us to visually compare different batches of data. On the one end it shows a summary of the data, while on the other end all data points are still visible. Thereby, it enables us to discuss individual interesting data points. Next to that, it gives an indication of the number of data points, which helps when comparing groups with a widely varying number of data points. An implementation was made in R that keeps the user in mind and supports fast usage in scenarios like comparing multiple data sources and displaying exponential data. The **beanplot** package is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=beanplot`.

Acknowledgments

This research received full support by the Dutch *Joint Academic and Commercial Quality Research & Development (Jacquard)* program on Software Engineering Research via contract 638.004.405 *EQUITY: Exploring Quantifiable Information Technology Yields.* Asymmetric beanplots were suggested by an anonymous referee.

References

Adler D (2005). ***vioplot**: Violin Plot.* R package version 0.2, URL `http://CRAN.R-project.org/package=vioplot`.

Bakker A, Biehler R, Konold C (2005). "Should Young Students Learn About Box Plots?" In G Burrill, M Camden (eds.), "Curricular Development in Statistics Education: International Association for Statistical Education 2004 Roundtable," pp. 163–173. International Statistical Institute, Voorburg, The Netherlands.

Box GEP, Hunter WG, Hunter JS (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Hoboken, NJ.

Chambers JM, Cleveland WS, Kleiner B, Tukey PA (1983). *Graphical Methods for Data Analysis.* Chapman & Hall, New York.

Hintze JL, Nelson RD (1998). "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician*, **52**(2), 181–184.

McGill R, Tukey JW, Larsen WA (1978). "Variations of Box Plots." *The American Statistician*, **32**(1), 12–16.

Neuwirth E (2007). ***RColorBrewer**: **ColorBrewer** Palettes.* R package version 1.0-2, URL `http://CRAN.R-project.org/package=RColorBrewer`.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

Sarkar D (2008). ***lattice**: Multivariate Data Visualization with R*. Springer-Verlag, New York.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

Affiliation:

Peter Kampstra
Faculty of Exact Sciences
VU University Amsterdam
De Boelelaan 1081a
NL-1081 HV Amsterdam, The Netherlands
E-mail: pkampst@cs.vu.nl
URL: http://www.cs.vu.nl/~pkampst/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/

Volume 28, Code Snippet 1 *Submitted:* 2008-09-19
November 2008 *Accepted:* 2008-10-28