Density Distribution Sunflower Plots

William D. Dupont*
and
W. Dale Plummer Jr.
Vanderbilt University School of Medicine

Abstract

Density distribution sunflower plots are used to display high-density bivariate data. They are useful for data where a conventional scatter plot is difficult to read due to overstriking of the plot symbol. The *x*-*y* plane is subdivided into a lattice of regular hexagonal bins of width *w* specified by the user. The user also specifies the values of *l*, *d*, and *k* that affect the plot as follows. Individual observations are plotted when there are less than *l* observations per bin as in a conventional scatter plot. Each bin with from *l* to *d* observations contains a light sunflower. Other bins contain a dark sunflower. In a light sunflower each petal represents one observation. In a dark sunflower, each petal represents *k* observations. (A dark sunflower with *p* petals represents between $pk - k/2$ and $pk + k/2$ observations.) The user can control the sizes and colors of the sunflowers. By selecting appropriate colors and sizes for the light and dark sunflowers, plots can be obtained that give both the overall sense of the data density distribution as well as the number of data points in any given region. The use of this graphic is illustrated with data from the Framingham Heart Study. A documented Stata program, called *sunflower*, is available to draw these graphs. It can be downloaded from the Statistical Software Components archive at http://ideas.repec.org/c/boc/bocode/s430201.html . (***Journal of Statistical Software*** **2003; 8 (3): 1–5.** Posted at http://www.jstatsoft.org/index.php?vol=8 .)

KEY WORDS: Scatter plot; Sunflower plot; Bivariate data; Density plot; Graphical statistics.

1 Introduction

The scatterplot is a powerful and ubiquitous graphic for displaying bivariate data [1]. These plots, however, become difficult to read when the density of points in a region becomes high (see Figure 1). Cleveland and McGill [2] introduced the sunflower plot as a solution to this problem. A sunflower is a number of short line segments, called petals, that radiate from a central point. In a sunflower plot, the *x*-*y* plane is divided into a lattice of regular square bins; a sunflower is placed in the center of each bin that contains one or more observations. They are drawn so that the number of petals of each sunflower equals the number of observations in the associated bin. Sunflower plots are effective at dealing with the overstrike problem that arises with high-density scatter plots. Unfortunately, information on the precise location of points is lost in low-density regions of the graph. This is particularly true when the bin size is large. Carr et al. [3] proposed plotting individual points at their exact location as long as there were less than four observations per bin. They also introduced hexagonal shaped bins that permit sunflowers to be more densely packed and that de-emphasize horizontal and vertical patterns that can be introduced by square bins. Scott [4] showed that hexagonal bins produce a lower integrated mean squared error for bivariate histograms than does any other bin shape that can tile the plane. Carr et al. [3] also experimented with using a hexagonal shaped symbol whose size increased monotonically as the number of observations in the associated bin increased. Huang et al. [5] introduced a similar graphic. These approaches give an excellent feel for the density distribution of the bivariate data. They do not, however, permit readers to estimate the number of observations in a given region. In addition, these graphs are not trivial to produce, and these authors have not provided software written in a common language that makes them easy to draw. In this paper we intro-

*From the Division of Biostatistics, S2323 Medical Center North, Vanderbilt University School of Medicine, Nashville, Tennessee 37232-2158. E-mail: william.dupont@vanderbilt.edu , dale.plummer@vanderbilt.edu .

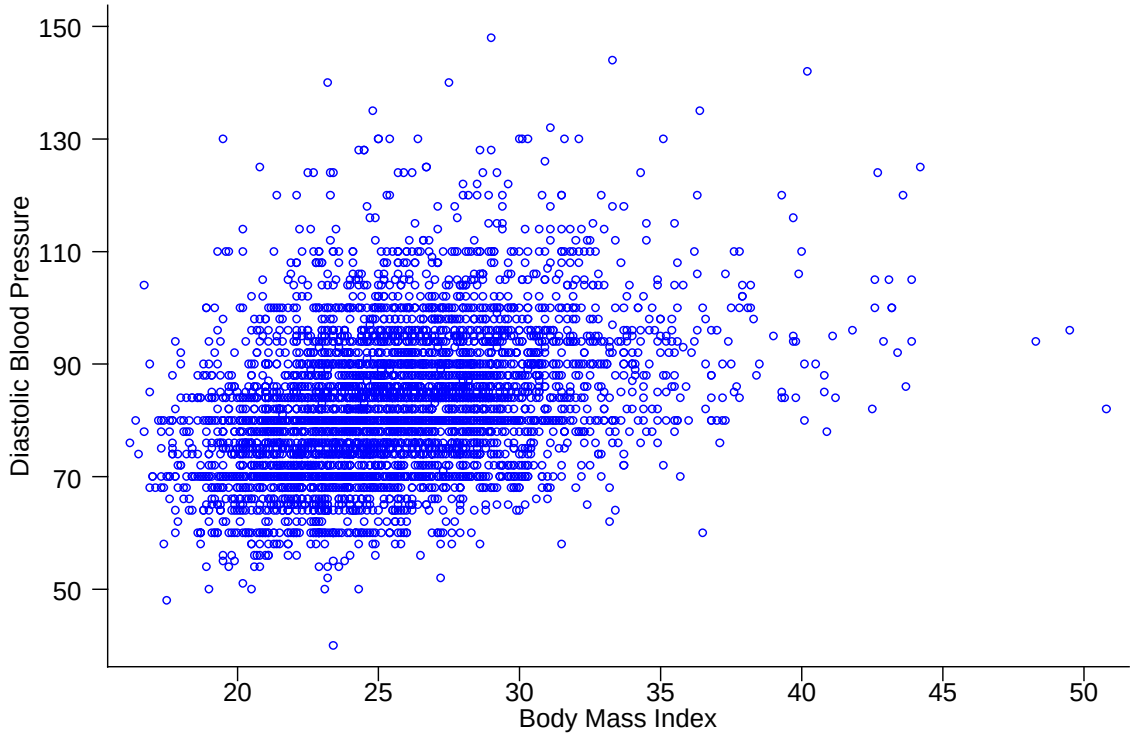

***Figure 1**: Scatter plot of the baseline diastolic blood pressure versus body mass index for 4689 subjects from the Framingham Heart Study [6,7]. Overstriking of many observations near the center of this graph makes it impossible to determine the density of observations for the most common values of these two variables.*

duce the density distribution sunflower plot. This graphic attempts to combine the best features of the sunflower plot and the density distribution graphics of Carr et al. [3] and Huang et al. [5]. A documented Stata program is available to draw these graphs.

2 Density Distribution Sunflower Plots

Figure 2 shows a density distribution sunflower plot of baseline diastolic blood pressure versus body mass index for subjects in the Framingham Heart Study [6,7]. This is the same data set displayed in Figure 1. Data points are represented in one of three ways: as small circles representing individual data points as in a conventional scatterplot, as light sunflowers, and as dark sunflowers. In a light sunflower each petal represents one observation. In Figure 2, light sunflowers are drawn in dark brown on a light green background. In a dark sunflower, each petal represents k observations, where k is specified by the user. (A dark sunflower with p petals represents between $pk - k/2$ and $pk + k/2$ observations.) In Figure 2, $k = 7$, and the dark sunflowers are drawn in black on a brown background. The first step in producing this graph is to define a lattice of hexagonal bins for the graph. The user specifies the bin width in the units of the x-axis. The bin height is then determined by the graphing software in such a way as to produce regular hexagonal bins. The user also specifies two thresholds l and d. Whenever there are less than l data points in a bin the individual data points are depicted at their exact location. When there are at least l but fewer than d data points in a bin they are depicted by a light sunflower. When there are at least d observations in a bin they

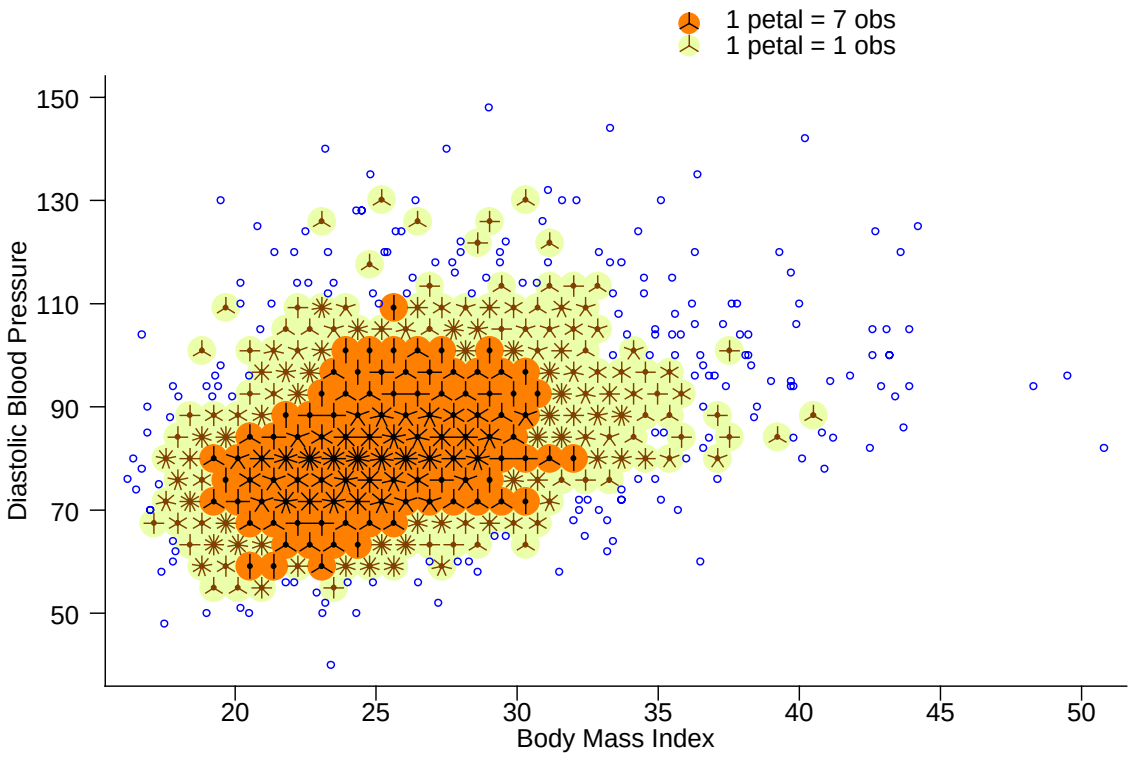

Figure 2: *A density distribution sunflower plot of the data from Figure 1. In this example, the x-y plane is divided into regular hexagonal bins of width 0.85 kg/m^2. Individual observations are depicted by blue circles at their exact location as long as there are less than 3 observations per bin. Observations in bins with higher densities are represented by light or dark sunflowers. Light sunflowers have green backgrounds and represent one observation for each petal. Dark sunflowers have brown backgrounds and represent 7 observations per petal. This plot conveys the density distribution of the observations while also allowing the reader to determine the number observations in any region with considerable precision.*

are depicted by a dark sunflower. If the number of observations in a bin is less than 1.5k but at least d then a dark sunflower is drawn as a single dot in the center of the bin. Similarly, if l = 1 and there is only one observation in a bin then a light sunflower is drawn as a single dot in the center of the bin. In Figure 2, l = 3 and d = 13. Note that the maximum density of observations represented by dark sunflowers in this figure is about 98 subjects per bin. The user can control the colors of the dark and light sunflowers, their background colors, the color used to depict individual data points, and the length and thickness of the lines used for light and dark sunflowers. Although color is helpful for these plots, black and white plots can be produced by drawing light sunflowers with black ink on a gray background and dark sunflowers with white ink on a black background. We have written a documented Stata program (*ado* file) to draw these plots, which is in the public domain [8]. It is based, in part, on public domain code authored by Steichen and Cox [9]. The user must have Stata Release 7 or a later version installed on her computer to use this program [10].

3 Discussion

The density distribution sunflower plot combines features of the original sunflower plot of Cleveland and McGill [2] with the graphics proposed by Carr et al. [3] and Huang et al. [5]. It shares with these latter graphics the ability to depict individual data points in low-density regions. If the bin size is kept small and

the background colors of light and dark sunflowers are chosen carefully, the density distribution sunflower plot does a good job at depicting the density distribution of the bivariate data. At this task it is comparable to the *Varebi* plots of Huang et al. [5] and the density plots depicted in Figures 8 and 9 of Carr et al. [3]. Our graphic also uses the hexagonal bins of Carr et al. [3]. Like the *Varebi* plots, our graphic can be redrawn interactively to account for changes in the ratio of the lengths of the *x*- and *y*-axes. An advantage of our approach is that it provides more information on the actual distribution of the data. The reader can determine the exact location of data points in low density regions, the exact number of data points in bins that contain light sunflowers, and can estimate to within $k/2$ observations the number of data points in bins with dark sunflowers. In contrast, the *Varebi* graphs and the area density graphs of Carr et al. [3] give only relative changes in the density of the data. An important advantage of our approach is that it may be easily implemented by users of an established statistical software package [10]. The density distribution sunflower plot could easily be extended to handle a wider range of density distributions by introducing more than two types of sunflowers (e.g. light, darker and darkest sunflowers). However, most high-density data sets that we have encountered can be effectively displayed using only light and dark sunflowers.

The density distribution sunflower plot is analogous to the stem-and-leaf plot of Tukey [11]. At a distance, stem-and-leaf plots look like histograms and provide a good intuitive depiction of the distribution of a univariate data set. However, the values of the individual data points can be determined from the plot by examining the individual values of the "leaves". Similarly, the density distribution sunflower plot can provide an intuitive picture of the bivariate distribution of two variables. Close inspection of the sunflowers, however, provides far more information about the actual data set than can be obtained from a conventional bivariate density plot.

Acknowledgment: This work was supported in part by NIH grants # R01 CA50468, 1 P30 CA68485 and 5 P30 DK26657. We thank Thomas J. Steichen and Nicholas J. Cox for making their software available [9]. We also thank Nicholas J. Cox for converting our Stata help file to SMCL and for some helpful edits, and the associate editor and his referees for their helpful suggestions. This paper used data supplied by the National Heart, Lung and Blood Institute, NIH, DHHS. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the National Heart, Lung and Blood Institute.

3 References

[1] Pagano, M. and Gauvreau, K. (2000), *Principles of Biostatistics* (2nd ed.), Pacific Grove, CA: Duxbury.

[2] Cleveland, W.S. and McGill, R. (1984), "The Many Faces of a Scatterplot," *Journal of the American Statistical Association*, **79**, 807-822.

[3] Carr, D.B., Littlefield, R.J., Nicholson, W.L., and Littlefield, J.S. (1987), "Scatterplot Matrix Techniques for Large N," *Journal of the American Statistical Association*, **82**, 424-436.

[4] Scott, D.W. (1988), "A Note on Choice of Bivariate Histogram Bin Shape," *Journal of Official Statistics*, **4**, 47-51.

[5] Huang, C., McDonald, J.A, and Stuetzle, W. (1997), "Variable Resolution Bivariate Plots," *Journal of Computational and Graphical Statistics*, **6**, 383-396.

[6] Framingham Heart Study (1997), *The Framingham Study – 40 Year Public Use Data Set*, Bethesda, MD: National Heart, Lung, and Blood Institute, NIH.

[7] Levy, D. (1999), *50 Years of Discovery: Medical Milestones from the National Heart, Lung, and Blood Institute's Framingham Heart Study*, Hackensack, NJ: Center for Bio-Medical Communication Inc.

[8] Dupont, W.D. and Plummer, W.D. Jr. (2002). "Sunflower: Stata Module to Draw Density Distribution Sunflower Plots," Stata program and help file downloadable from http://ideas.repec.org/c/boc/bocode/s430201.html. Accessed December 18, 2002.

[9] Steichen, T.J. and Cox, N.J. (1999). "Flower: Stata Module to Draw Sunflower Plots," Stata program and help file downloadable from http://ideas.repec.org/c/boc/bocode/s393001.html. Accessed December 6, 2002.

[10] StataCorp. (2001), *Stata Statistical Software: Release 7.0*, College Station, TX: Stata Corporation.

[11] Tukey, J. (1977). *Exploratory Data Analysis*. Reading MA: Addison-Wesley