The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

The Stata Journal (2005)
**5**, Number 3, pp. 371–384

# Using density-distribution sunflower plots to explore bivariate relationships in dense data

William D. Dupont
william.dupont@vanderbilt.edu

W. Dale Plummer, Jr.
dale.plummer@vanderbilt.edu

Department of Biostatistics
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 scatterplot is difficult to read due to overstriking of the plot symbol. The $x$–$y$ plane is subdivided into a lattice of small, regular, hexagonal bins. These bins are divided into low-, medium-, and high-density groups. In low-density bins, the individual observations are plotted as in a conventional scatterplot. Medium- and high-density bins contain light and dark sunflowers, respectively. In a light sunflower, each petal represents one observation. In a dark sunflower, each petal represents a specific number of 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.

Sunflower plots are also contrasted with contour plots of bivariate kernel-density estimates. The appearance of these plots is markedly affected by the choice of smoothing parameters and the spacing of points at which the probability density function is evaluated. Sunflower plots can be helpful in guiding the selection of these parameters and in distinguishing between chance and systematic variation in the distribution of bivariate data.

**Keywords:** gr0016, scatterplot, sunflower plot, bivariate data, density plot, probability density function, graphical statistics

## 1 Introduction

The scatterplot is a powerful and ubiquitous graphic for displaying bivariate data. These plots, however, become difficult to read when the density of points in a region becomes high (see figure 1).

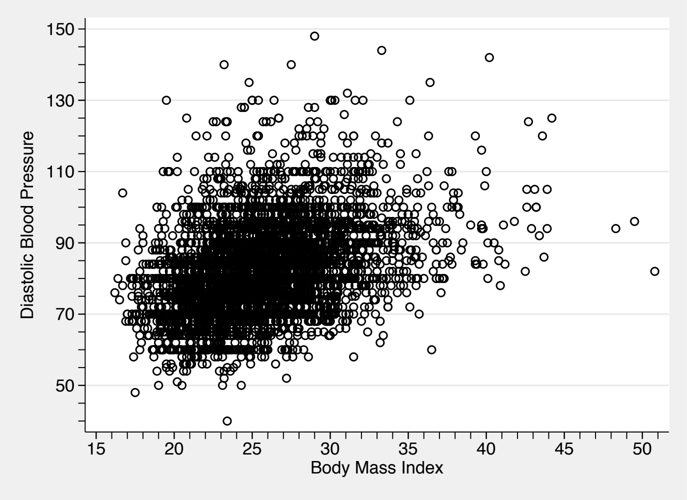

Figure 1: Scatter plot of the baseline diastolic blood pressure versus body mass index for 4,689 subjects from the Framingham Heart Study (1997). Overstriking of many observations near the center of this graph makes it impossible to determine the density of observations for the most common values of these two variables.

## 1.1 Bivariate kernel-density estimation contour plots

One approach to this problem is to directly estimate the bivariate probability density function (pdf) from a sample $\{(x_i, y_i) : i = 1, \ldots, n\}$ using a kernel density estimator (Wand and Jones [1995]). The shape of this pdf can then be graphed using a contour plot. We estimate the pdf $f(x, y)$ by $\widehat{f}\,(x, y;\ h_x, h_y)$, which is a weighted sum of the observations near $(x, y)$. The weights are a function of the distance between $(x_i, y_i)$ and $(x, y)$; the bandwidths, $h_x$ and $h_y$, which control the degree of smoothing in the $x$ and $y$ direction; and the kernel, which is itself a known pdf. We evaluate $\widehat{f}\,(x, y;\ h_x, h_y)$ at a rectangular array of evenly spaced points at $N_x$ values of $x$ and $N_y$ values of $y$. The appearance of the resulting contour plot is affected by $h_x$, $h_y$, $N_x$, $N_y$, and the kernel. Figure 2 shows six contour plots of the data in figure 1 using different values of these parameters and an independent standard normal kernel. Note that the shape of these plots is dramatically affected by these parameters and that the degree of skewness of the blood-pressure distribution is markedly affected by the smoothing parameter $h_y$ and the grid density, which is determined by $N_x$ and $N_y$. A problem with these plots is that they are visually detached from the sample that generated them. This makes it impossible to tell whether the complex estimated pdf in the lower-right panel results from overfitting or if the simple estimate in the upper-right panel results from excessive smoothing. Our intent is to produce a graphic that, like the contour plots in figure 2, gives a sense of the underlying pdf, but which, like a scatterplot for low-density data gives a sense of the size of the sample and of fluctuations in data density that are due to both systematic and chance variation.

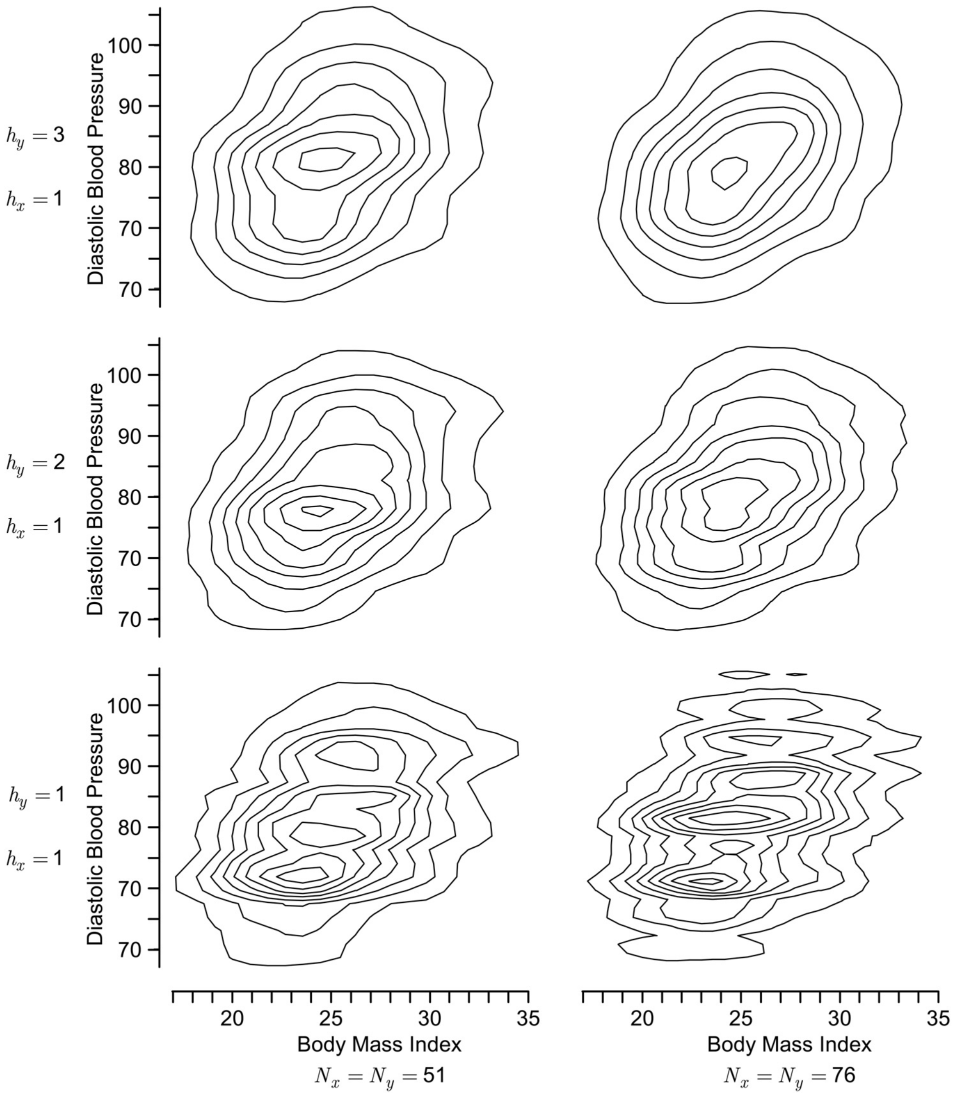

Figure 2: Bivariate kernel-density estimation contour plots derived from the data displayed in figure 1. These plots are drawn with different values of the parameters $h_x$, $h_y$, $N_x$, and $N_y$. Note that the appearance of these pdf estimates vary markedly with the values of these parameters. It is difficult to tell from these graphs the extent to which the blood-pressure data is skewed or to distinguish between true features of the underlying distribution and chance variation. These graphs were drawn using the R functions `bkde2D` and `contour` from the KernSmooth and graphics packages, respectively (see Wand and Ripley [2005] and R Development Core Team [2004]).

## 1.2 Density-distribution sunflower plots

In a density-distribution sunflower plot, 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 fewer than $l$ observations per bin as in a conventional scatterplot. Each bin with $l$ 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.

Figure 3 shows such a plot of baseline diastolic blood pressure (*yvar*) versus body mass index (*xvar*) for subjects in the Framingham Heart Study (1997) (see also Levy [1999]). This is the same dataset displayed in figure 1. In figure 3, light sunflowers are colored gray on a light gray background. Dark sunflowers are colored black on a dark gray background. Each petal of the dark sunflowers represents $k = 5$ subjects. The first step in producing this graph is to define a lattice of hexagonal bins for the graph. The user specifies the bin width $w$ 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 can also specify the two thresholds, $l$ and $d$. In figure 3, $l = 3$ and $d = 13$, which are the default values for these parameters. The default petal weight, $k$, is chosen so that the maximum number of dark sunflower petals equals 14. Note that the maximum density of observations represented by dark sunflowers in this figure is about 60 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 also be effective, as shown in figure 3. In this figure, the bin width $w$ was selected to equal 0.9 kg/m$^2$. The bin width functions as a smoothing parameter in much the same way as $h_x$ and $h_y$ in bivariate kernel-density estimation contour plots. Larger bin widths provide greater smoothing. In general, we recommend using narrow bin widths, as this reveals more information about the data sample and allows the reader to use her judgment in assessing the extent to which variation in data density is due to chance fluctuation or to the underlying shape of the pdf that generated the sample. For example, in figure 3, the overall shape of this underlying distribution is clear, even though there is fluctuation in the data density of adjacent bins that we would ascribe to chance. Even if one's ultimate objective is to estimate the pdf using a contour plot as in section 1.1, plotting sunflower plots can be helpful in guiding the selection of $h_x$, $h_y$, $N_x$, and $N_y$.

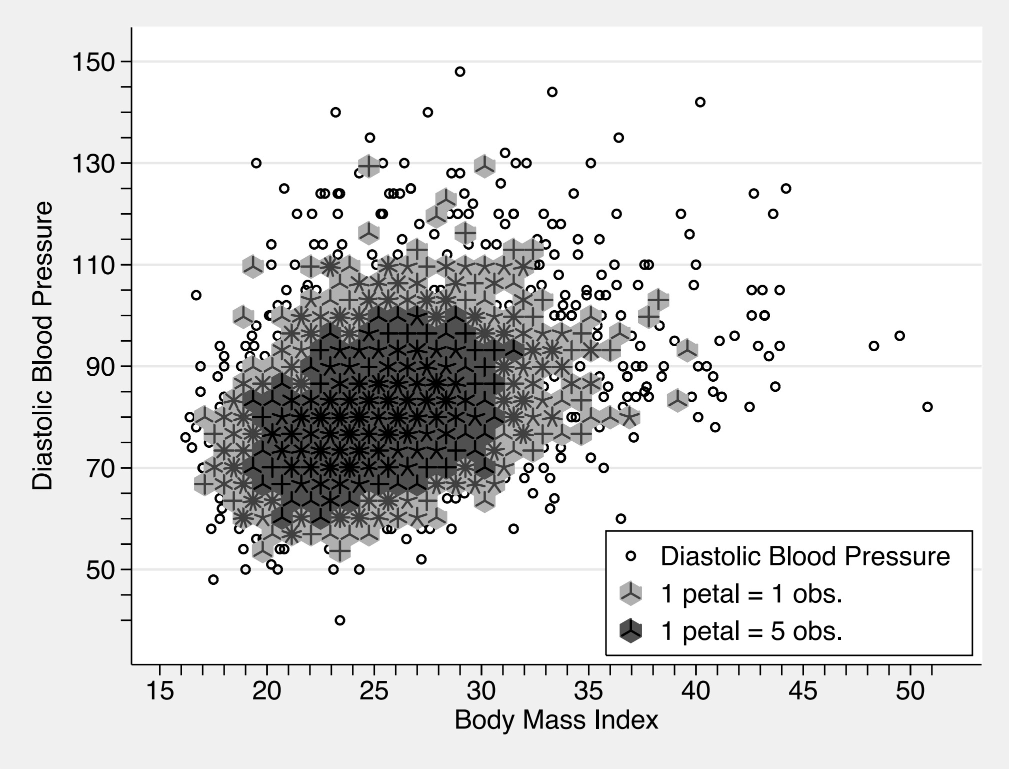

Figure 3: A density-distribution sunflower plot of the data from figure 1. In this example, the $x$–$y$ plane is divided into regular hexagonal bins of width 0.9 kg/m$^2$. Individual observations are depicted by black circles at their exact location as long as there are fewer than 3 observations per bin. Observations in bins with higher densities are represented by light or dark sunflowers. Light sunflowers are gray with light gray backgrounds and represent one observation for each petal. Dark sunflowers are black with dark gray backgrounds and represent 5 observations per petal. This plot conveys the density distribution of the observations while also allowing the reader to determine the number of observations in any region with considerable precision.

## 1.3 Overview of the literature

Sunflower plots were first introduced by Cleveland and McGill (1984), who suggested plotting light sunflowers in rectangular bins. A disadvantage of these plots was that 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. It can also be difficult to discern individual petals in high-density regions. Carr et al. (1987) proposed plotting individual points at their exact location as long as there were fewer 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 (1988) 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. (1987) also experimented with using a hexagonal-shaped symbol whose size increased monotonically as the number of observations in the associated bin increased. Huang, McDonald, and Stuetzle (1997) introduced a similar graphic called a *Varebi* plot. 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. We introduced the density-distribution sunflower plot (Dupont and Plummer, Jr. [2003]), which attempts to combine the best features of the original sunflower plot and the density-distribution graphics of Carr et al. (1987) and Huang, McDonald, and Stuetzle (1997). A program to produce these plots has been available in Stata starting with version 8.1; see [R] **sunflower**.

## 2 Example: Comparing the distribution of diastolic blood pressure and body mass index in men and women

It should be noted that by simply giving the command `sunflower yvar xvar`, we can generate an informative sunflower graph. In this section, we illustrate how one might generate a publication quality graphic using some of the powerful `twoway` options that also apply to the `sunflower` command. We explore the relationship between diastolic blood pressure and body mass index in men and women from the Framingham Heart Study (1997). The dataset used in this example contains variables `dbp`, `bmi`, and `male`, which give the patient's baseline blood pressure, body mass index, and gender, respectively (`male` is true for men). Figure 4 presents sunflower plots of `dbp` against `bmi` for women and men. We could have produced this graph using the `by(male)` option. We have not done this primarily because the `by()` option forces the legend to be placed outside of the axes, which reduces the information content per square centimeter of graph. Instead, we used two separate commands to produce the top and bottom panels of figure 4, which were concatenated later.

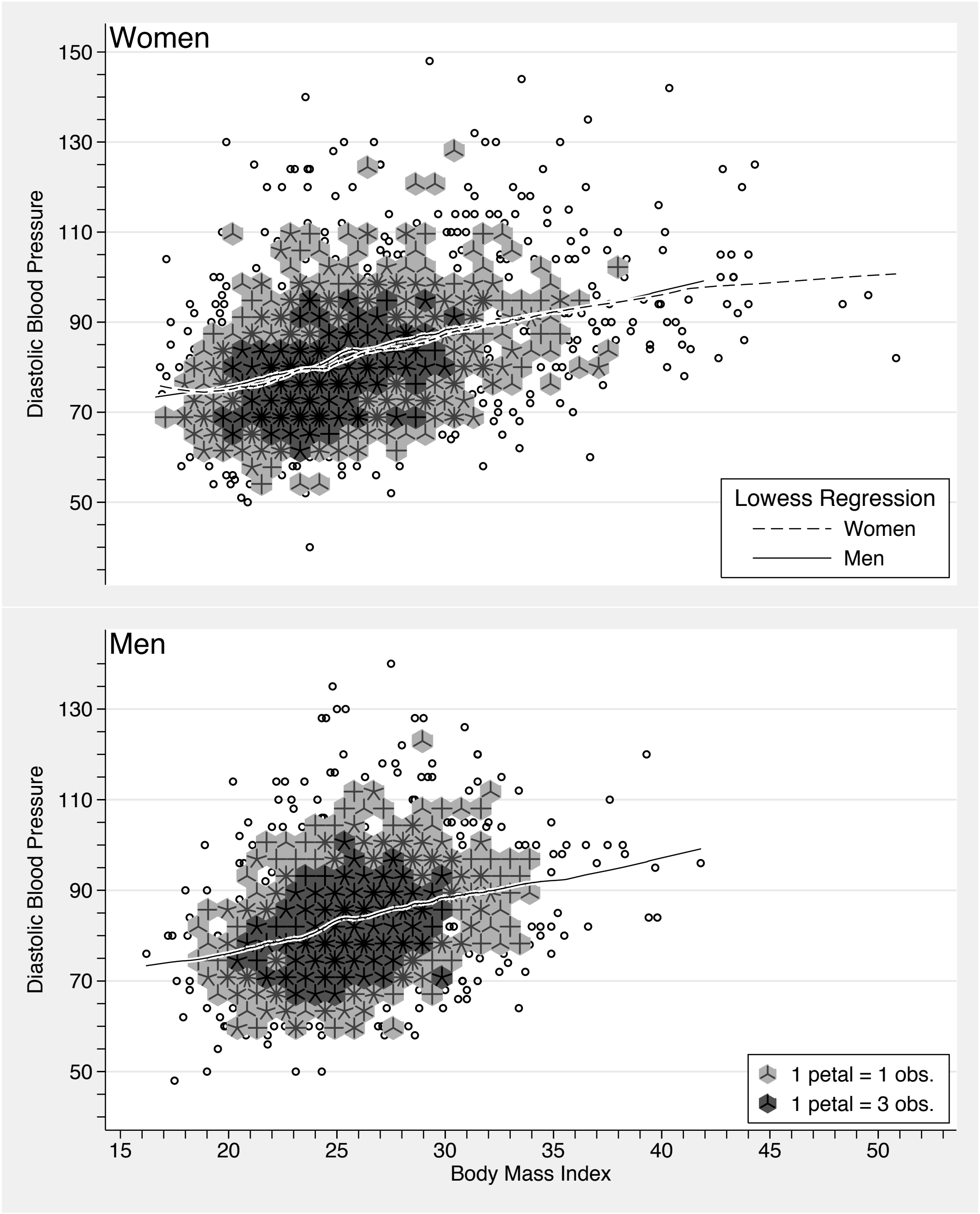

Figure 4: Sunflower plots for men and women from the Framingham Heart Study (1997). Lowess regression curves for each gender are also shown. Although the regression curves are very similar, these plots reveal that the distribution of diastolic blood pressure is more skewed in women than in men.

The top panel of figure 4 was generated as follows:

```
. set scheme sj

. sunflower dbp bmi if ~male, binwidth(.9) lflcolor(gs4) dbcolor(gs5)
>   msymbol(oh) petalweight(3) ylabel(50 (20) 150, angle(0)) ytick(35 (5) 145)
>   plot(lowess dbp bmi if male, bwidth(.2)
>       clcolor(white) clwidth(thick) clpattern(solid)
>     || lowess dbp bmi if  male, bwidth(.2) clwidth(thin) clpattern(solid)
>     || lowess dbp bmi if ~male, bwidth(.2) clc(white) clw(thick) clp(solid)
>     || lowess dbp bmi if ~male, bwidth(.2) clwidth(thin) clpattern(dash))
>   xlabel(15 50) xtick(15 53) xscale(off) ytitle(Diastolic Blood Pressure)
>   title("Women", position(11) ring(0) )
>   legend(position(5) ring(0) cols(1) subtitle("Lowess Regression")
>     order(7 "Women" 5 "Men")) xsize(5.25) ysize(3.25)
Bin width          =        .9
Bin height         =   4.93942
Bin aspect ratio   =   4.75296
Max obs in a bin   =        47
Light              =         3
Dark               =        13
X-center           =      24.5
Y-center           =        80
Petal weight       =         3
------------------------------------------------------------------
     flower      petal      No. of      No. of  estimated      actual
       type     weight      petals     flowers       obs.        obs.
------------------------------------------------------------------

       none                                          218         218
      light          1           3          36       108         108
      light          1           4          27       108         108
      light          1           5          17        85          85
      light          1           6          14        84          84
      light          1           7          10        70          70
      light          1           8          10        80          80
      light          1           9           7        63          63
      light          1          10           5        50          50
      light          1          11          11       121         121
      light          1          12           6        72          72
       dark          3           4           3        36          39
       dark          3           5          14       210         209
       dark          3           6          11       198         195
       dark          3           7          10       210         209
       dark          3           8          12       288         295
       dark          3           9           4       108         108
       dark          3          10           4       120         121
       dark          3          11           4       132         130
       dark          3          12           1        36          36
       dark          3          13           5       195         194
       dark          3          16           1        48          47
------------------------------------------------------------------
                                                    2640        2642
```

This `sunflower` command produces a graph of `dbp` against `bmi` among women (patients for whom `male` is false). We have chosen a bin width of 0.9 kg/m$^2$. In general, we recommend the smallest bin width for which individual petals are readily discernable. This is because larger bin sizes reduce the amount of information presented about the observed data. For monochrome graphs, we recommend a lighter gray for the light sunflowers and a darker gray for the dark sunflower backgrounds than the default values. This is done with the `lflcolor(gs4)` and `dbcolor(gs5)` options, respectively, and produces graphs that enhance the contrast between light and dark sunflowers. We explicitly set the petal weight for dark sunflowers to equal three (`petalweight(3)`) in the top panel in order to force these weights to be the same for men and women. The other options used in this command are standard `twoway` options.

We have superimposed on this graph lowess regression curves of blood pressure against body mass index for women and men (Cleveland [1993]). The background for these curves varies from almost black to white. To make the curves visible regardless of the background, we plot each curve twice with a thin black line superimposed over a thick white line. The first three keys in the figure legend of a `sunflower` plot are the plot symbol for individual observations, the light sunflower symbol, and the dark sunflower symbol. In the upper panel, we have suppressed these keys because they are given in the lower panel. Subsequent keys are numbered in the same order that they are specified in the option list. Thus keys four and six are for thick white lines that enhance the visibility of the regression lines in the center of the graph. The black solid line for men is the fifth key, and the dashed line for women is the seventh key. The `legend()` option that displays these keys is `order(7 "Women" 5 "Men")`. The $x$-axis is suppressed in the top panel with the `xscale(off)` option because we will draw this axis in the lower panel.

Figure 4 could have been generated using the `graph combine` command. Unfortunately, it is very difficult to do this without changing the aspect ratio of at least one of the graph panels. For this reason, we do not recommend using this command if maintaining regular bin shapes is desired. Instead, we prefer to merge the two panels together with a graphics editor or to import them as separate figures into a LaTeX document, as was done here. Note that in generating the top panel, we have included `xlabel` and `xtick` options in spite of having suppressed the $x$-axis using the `xscale(off)` option. We did this in order to ensure that $x$-axes of the top and bottom panels lined up correctly. We recommend that if you use the `xscale(off)` option in this way you also do a test plot with the `xscale(off)` option removed to confirm that the $x$-axis is really aligned the way you want it. (We were unable to correctly align the top and bottom panels of this figure using the `range` suboption of the `xscale()` option.)

The output from this command gives the bin dimensions in units of the $x$ and $y$ variables and other information. In this example, it specifies that the minimum number of observations required for a light or dark sunflower is 3 and 14, respectively. One of the bins is centered at the median values of the $x$ and $y$ variables, which is at 24.5 kg/m$^2$ and 80 mm Hg, respectively. The other bin centers are a function of this location and the bin width and height. The petal weight for dark sunflowers is three. A table listing the number of distinct observations and the number of light and dark sunflowers with the specified petal numbers is also given. This table can be helpful in conducting a detailed study of the distribution of the study dataset.

The bottom panel of figure 4 was generated by the following code. The `order()` suboption of the `legend()` option generates the key for light and dark sunflowers, respectively. This plot was restricted to men by the `keep if male` statement. We could also have done this by adding an `if male` qualifier to the `sunflower` command. However, the range of blood pressures is smaller for men than for women. Deleting the records of women before giving the `sunflower` command causes the lower panel of figure 4 to have a smaller range for the $y$ variable.

```
. keep if male
(2642 observations deleted)
. sunflower dbp bmi , binwidth(.9) lflcolor(gs4) dbcolor(gs5) msymbol(oh)
>   plot(lowess dbp bmi, bwidth(.2) clcolor(white) clwi(thick) clpa(solid)
>     || lowess dbp bmi, bwidth(.2) clwidth(thin) clpattern(solid))
>   xlabel(15 (5) 50) xtick(16 (1) 53) ytitle(Diastolic Blood Pressure)
>   ylabel(50 (20) 130, angle(0)) ytick(40 (5) 140)
>   title("Men", position(11) ring(0))
>   legend(position(5) ring(0) cols(1) order(2 3)) xsize(5.25) ysize(3.25)
  (output omitted)
```

Finally, the upper and lower panels of figure 4 were concatenated together using LaTeX. This could also have been done by cutting and pasting into a graphic editor, such as PowerPoint or PhotoShop.

# 3 Assigning observations to bins

To draw a density-distribution sunflower plot, we first need to tile the $x$–$y$ plane with regular hexagonal bins and determine the number of observations that fall in each bin. We do this by determining the bin width $d$ in millimeters—that is, in units of physical distance on the graph, as opposed to units of the $x$ or $y$ variables. This value is used to produce two rectangular tilings of the plane, which we will refer to as the odd and even tilings (see figure 5). Each tile is $d$ mm wide and $d\sqrt{3}$ mm high. The odd tiles are positioned in such a way that their corners are at the centers of even tiles and vice versa. Points on the tile boundaries are randomly assigned to one of the adjacent tiles. Then each point on the $x$–$y$ plane lies in exactly one odd and one even rectangular tile. It is easily shown that the plane is also tiled by regular hexagonal bins of width $d$ whose centers lie at the centers of the odd and even rectangular tiles and that an observation lies in a hexagonal bin if it is also in the rectangular tile with the same center and is closer to that tile center than it is to any other rectangular tile center. For example, in figure 5 the observation $(x, y)$ lies in both the odd and even rectangular tiles on the left. It is closer to its even tile center than its odd tile center, which implies that it lies in the even hexagonal bin on the lower left of this figure. We assign each observation to a unique odd and even rectangular tile, determine its distance from the centers of these two tiles, assign it to the hexagonal bin with the closest bin center, and then count the number of observations in each hexagonal bin.

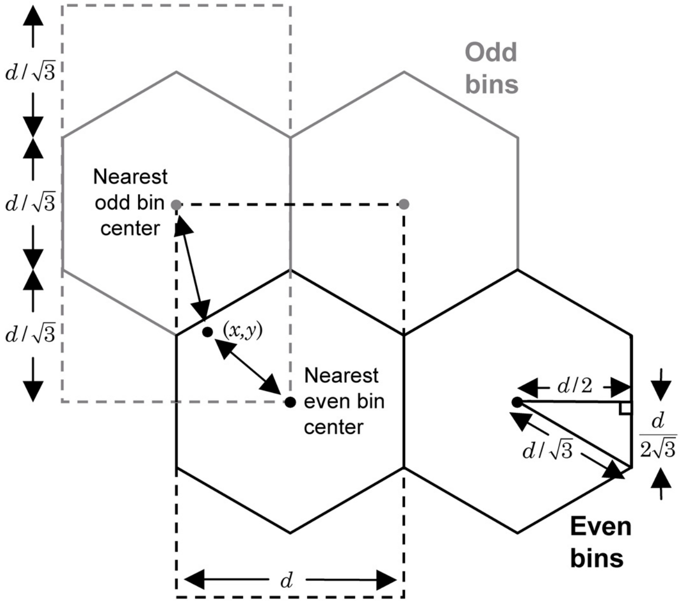

Figure 5: To draw a density-distribution sunflower plot, we first need to determine the number of observations in each hexagonal bin. To do this, we tile the $x$–$y$ plane with rectangular bins of width $d$ and height $d\sqrt{3}$, as shown above. Two such overlapping tilings are generated that we call odd (gray) and even (black). We determine the odd and even bin that contains each observation and then determine which of these two bins has the bin center that is closest to the observation. This observation lies in the hexagonal bin of width $d$ that is centered on this nearest rectangular bin center. For example, the observation $(x, y)$ in this figure lies in both the odd and even rectangular bins on the left. It is closer to the center of the lower of these two bins, which implies that it is in the lower-left hexagonal bin in the figure. To make the hexagonal bins regular, distance measured in terms of the $x$ and $y$ variables must be converted to physical distance on the graph (see text).

From a programmer's viewpoint, the greatest challenge in drawing these plots is that we want the bin shapes to be regular. The value of $d$ is initially specified in units of the $x$ variable. To assign observations to correctly shaped bins, we must convert $d$ from units of $x$ to millimeters and then convert this distance into units of $y$. This requires determining the aspect ratio of bin heights to widths in units of $y$ and $x$, which is a complex function of the range of the $x$ and $y$ variables; the graph aspect ratio; and the labels, titles, and legends, which affect the physical lengths of the $x$- and $y$-axes. Fortunately, Stata possesses functions that allows us to map observations expressed in terms of $x$ and $y$ variables into physical locations on the graph and to derive the inverse of this mapping. This permits us to construct these graphs correctly.

# 4 Using big bins

Large sunflowers are more easily seen at a distance by large audiences. Large bins also have a smoothing effect in that they tend to average out chance variation in the sample distribution. The disadvantages of large bins are that the subjective sense of the pdf is reduced and information about the sample distribution is lost. The following code generated figure 6, which uses a bin width of 2 kg/m$^2$:

```
. sunflower dbp bmi, binwidth(2) dflwidth(medthick) lflcolor(gs4) dbcolor(gs5)
>   xlabel(15 (5) 50) xtick(16 (1) 53) ylabel(50 (20) 150, angle(0))
>   ytick(40 (5) 145) legend(position(5) ring(0) order(2 3))
  (output omitted)
```

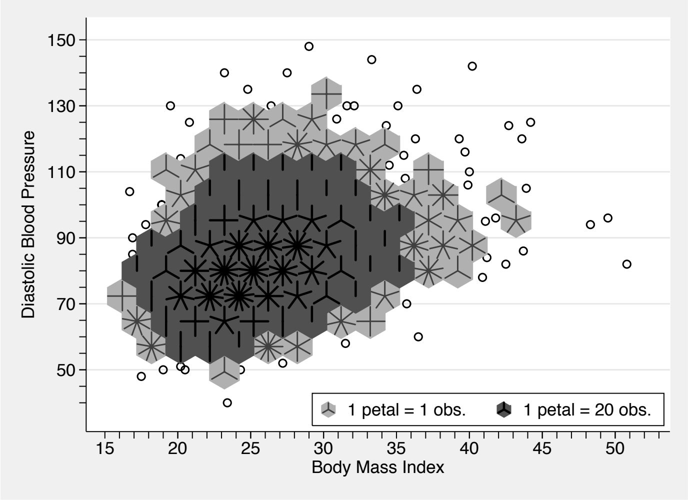

Figure 6: This graph was generated from the same data as in figure 3 but uses larger bins and thicker dark sunflower petals. This figure is easier to see at a distance. The large bins smooth the data but reduce the amount of information displayed. The subjective sense of the data's pdf is reduced.

# 5 Discussion

The density-distribution sunflower plot combines features of the original sunflower plot of Cleveland and McGill (1984) with the graphics proposed by Carr et al. (1987) and Huang, McDonald, and Stuetzle (1997). 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 of depicting the density distribution of the bivariate data. At this task, it is comparable to the *Varebi* plots of Huang, McDonald, and Stuetzle (1997) and the density plots depicted in figures 8 and 9 of Carr et al. (1987). Our graphic also uses the hexagonal bins of Carr et al. (1987).

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 and 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. (1987) give only relative changes in the density of the data. An important advantage of our approach is that it may be easily implemented by Stata users. 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 datasets that we have encountered can be effectively displayed using only light and dark sunflowers.

Bivariate kernel-density estimation contour plots can be effective for displaying estimates of the pdf (Wand and Jones [1995]). These estimates are strongly affected by the choice of smoothing parameters and the grid density at which the pdf is estimated. Sunflower plots can provide useful guidance in choosing the values of these parameters.

The density-distribution sunflower plot is analogous to the stem-and-leaf plot of Tukey (1977). At a distance, stem-and-leaf plots look like histograms and provide a good intuitive depiction of the distribution of a univariate dataset. 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 the other types of bivariate density plots discussed in this paper.

## 6 Acknowledgments

We thank Thomas J. Steichen and Nicholas J. Cox for making their software available (Steichen and Cox 1999), William W. Gould for incorporating density-distribution sunflower plots into Stata, and Jeffrey S. Pitblado for writing the version 8 edition of this program. 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. This work was supported in part by the Cancer Center Support Grant P30-CA68485 and the Clinical Nutrition Research Unit Core Grant P30 DK26657 from the National Institutes of Health.

## 7 References

Carr, D. B., R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. 1987. Scatterplot matrix techniques for large N. *Journal of the American Statistical Association* 82: 424–436.

Cleveland, W. S. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.

Cleveland, W. S. and R. McGill. 1984. The many faces of a scatterplot. *Journal of the American Statistical Association* 79: 807–822.

Dupont, W. D. and W. D. Plummer, Jr. 2003. Density distribution sunflower plots. *Journal of Statistical Software* 8(3): 1–5.

Framingham Heart Study. 1997. *The Framingham Study—40-year public use dataset*. Bethesda, MD: National Heart, Lung, and Blood Institute, NIH.

Huang, C., J. A. McDonald, and W. Stuetzle. 1997. Variable resolution bivariate plots. *Journal of Computational and Graphical Statistics* 6: 383–396.

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.

R Development Core Team. 2004. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
*http://www.R-project.org.*

Scott, D. W. 1988. A note on choice of bivariate histogram bin shape. *Journal of Official Statistics* 4: 47–51.

Steichen, T. J. and N. J. Cox. 1999. flower: Stata module to draw sunflower plots. *http://ideas.repec.org/c/boc/bocode/s393001.html.*

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison–Wesley.

Wand, M. P. and M. C. Jones. 1995. *Kernel Smoothing*. London: Chapman & Hall.

Wand, M. P. and B. D. Ripley. 2005. KernSmooth 2.22. R package.
*http://web.maths.unsw.edu.au/~wand.*

## About the Authors

William D. Dupont is a professor of Biostatistics and Preventive Medicine at Vanderbilt University School of Medicine. His interests include the epidemiology of benign breast disease, power and sample size calculations, statistical graphics, and teaching intermediate level biostatistics to physician scientists. He is the author of *Statistical Modeling for Biomedical Researchers* (Cambridge University Press, 2002), which uses Stata to teach biostatistics to this audience.

W. Dale Plummer, Jr., is a systems analyst in the Department of Biostatistics at Vanderbilt University School of Medicine. He wrote the original Version 7 edition of the `sunflower` program. His other contributions include writing the computer code for *PS*, a statistical freeware package for power and sample size calculations
(*http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize*).