Sunflower Plot Statlet

statgraphics 18® centurion

Revised: 10/9/2017

Summary

The *Sunflower Plot Statlet* is used to display an X-Y scatterplot when the number of observations is large. To avoid the problem of overplotting point symbols with large amounts of data, glyphs in the shape of sunflowers are used to display the number of observations in small regions of the X-Y space. The original idea for the sunflower plot dates back to the article by Cleveland and McGill (1984).

Sample StatFolio: *sunflowerstatlet.sgp*

Sample Data

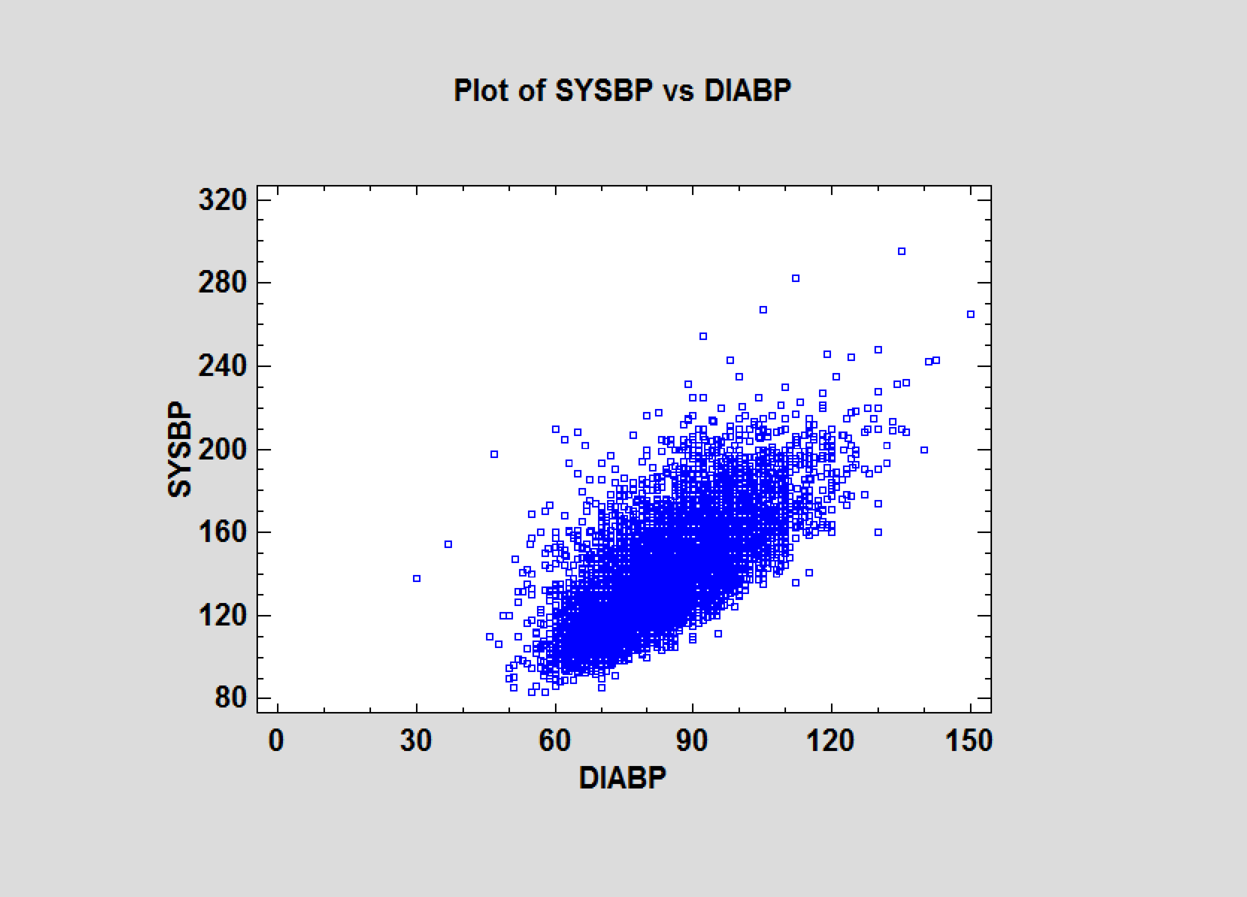
The file *Framingham heart study.sgd* contains data from a long-term study of cardiovascular disease performed in Framingham, MA. The file contains 11,627 records with information from 4,434 patients between 1956 and 1968.

The first several rows and columns of that file are shown below:

RANDID	SEX	TOTCHOL	AGE	SYSBP	DIABP
2448	1	195	39	106	70
2448	1	209	52	121	66
6238	2	250	46	121	81
6238	2	260	52	105	69.5
6238	2	237	58	108	66
9428	1	245	48	127.5	80
9428	1	283	54	141	89
10552	2	225	61	150	95
10552	2	232	67	183	109
2448	1	195	39	106	70

Motivation

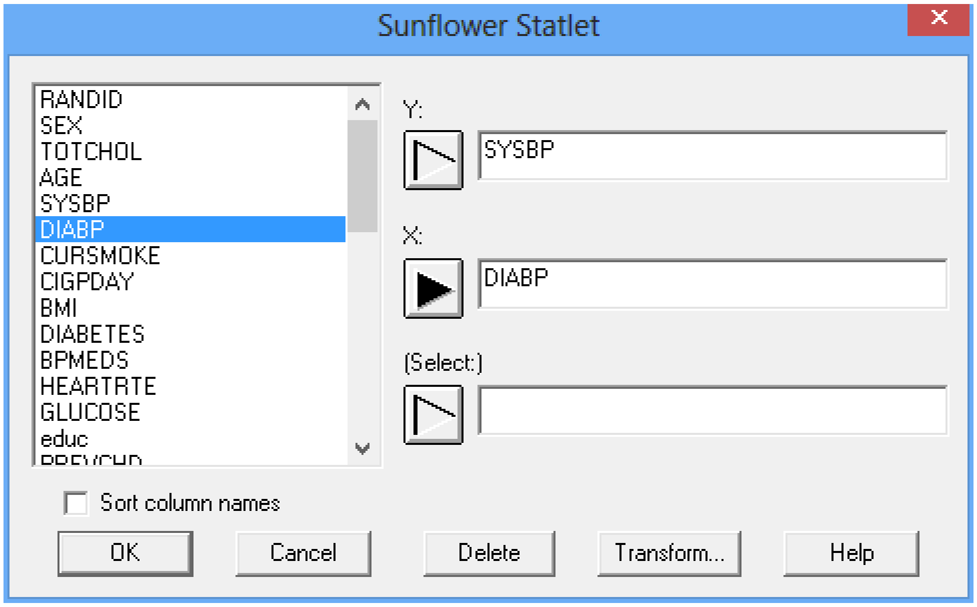
Suppose we wanted to display the relationship between systolic blood pressure and diastolic blood pressure using all of the data in the sample data file. A natural plot to select is the X-Y Scatterplot, which produces the following graph:



While we can sense the positive correlation between the two variables, there is much overplotting since the data consist of 11,627 observations. The *Sunflower Plot* is designed to extract more information from the data by indicating the number of points at different locations on the graph.

Data Input

The data input dialog box requests the names of the columns containing the data values to be plotted:



Y: SYSBP

X: DIABP

(Select:)

☐ Sort column names

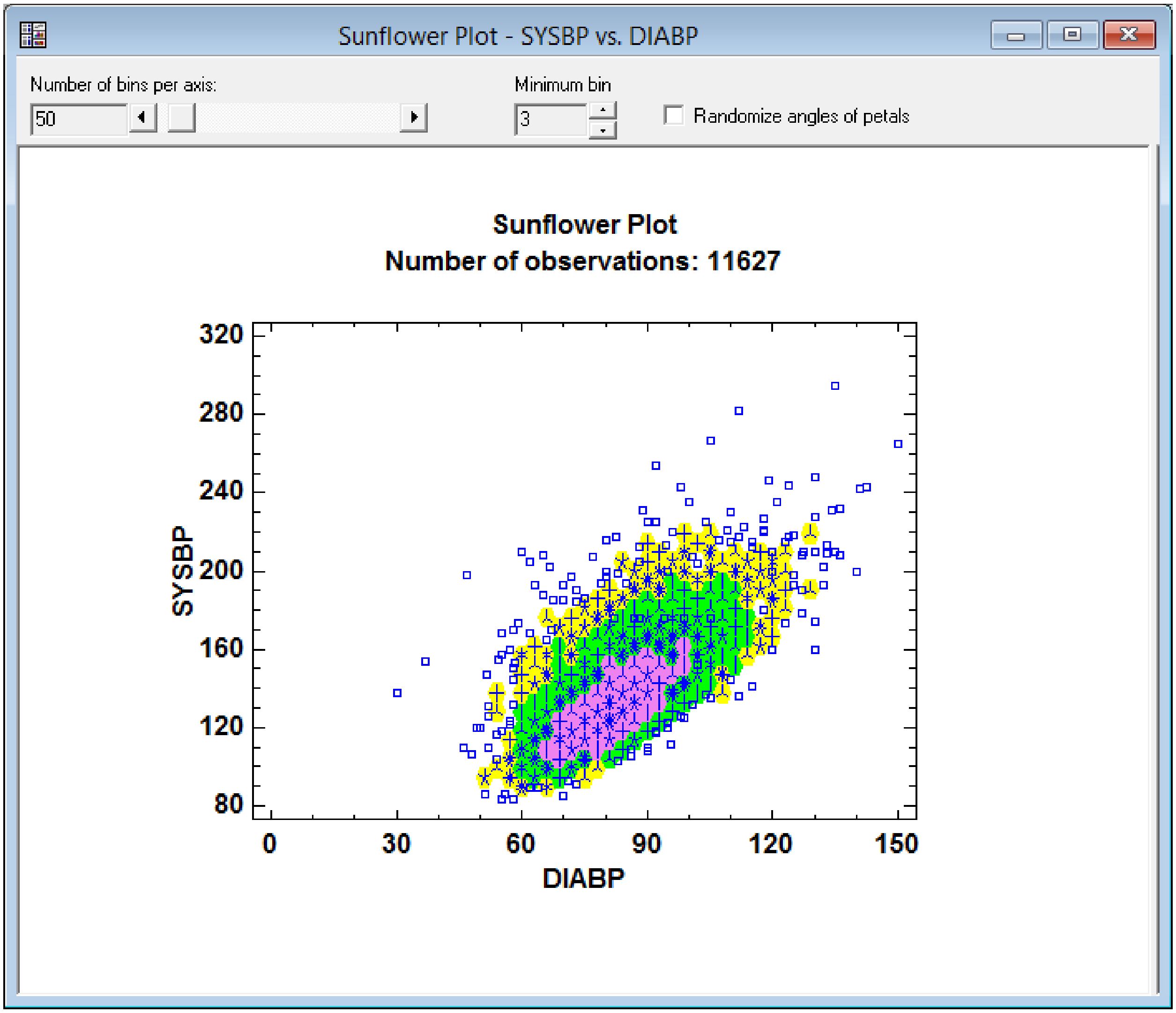
- **Y:** name of a numeric column containing the values to be plotted on the vertical axis.

- **X:** name of a second numeric column containing the values to be plotted on the horizontal axis.

- **Select:** optional subset selection.

Statlet

The output of this procedure is displayed in a dynamic Statlet window.



Sunflower Plot - SYSBP vs. DIABP

Number of bins per axis: 50

Minimum bin 3

□ Randomize angles of petals

The X-Y space is first divided into a *b*-by-*b* grid of non-overlapping hexagons, where *b* is the *Number of bins per axis* specified on the Statlet toolbar. The number of observations contained in each hexagon is then calculated. The graph is created as follows:

1. If a hexagonal region contains less than *m* observations, a separate point symbol is drawn for each observation. By default, *m* = 3 but may be changed using the *Minimum bin frequency* control on the Statlet toolbar.

2. If a hexagonal region contains between *m* and 10 observations, the hexagonal region is drawn using fill color #1 and one ray is drawn from the center of the hexagon to the exterior for each observation in that region. In the graph above, a yellow hexagon with 4 rays indicates a region containing 4 observations.

3. If a hexagonal region contains between 11 and 94 observations, the hexagonal region is drawn using fill color #2. In this region, each ray represents 7 observations. (The number of observations in the region is rounded to the nearest multiple of 7.) In the graph above, a green hexagon with 4 rays indicates a region containing 28 observations.

4. If a hexagonal region contains between 95 and 658 observations, the hexagonal region is drawn using fill color #3. In this region, each ray represents 49 observations. (The number of observations in the region is rounded to the nearest multiple of 49.) In the graph above, a magenta hexagon with 4 rays indicates a region containing 196 observations.

5. If a hexagonal region contains 659 or more observations, the hexagonal region is drawn using fill color #4. Each ray then represents 343 observations.

The sunflower plot shows clearly that the points are densest in the region around X=80 and Y=120.

Statlet Controls

- **Number of bins per axis**: the number of hexagons *b* filling the space along each axis.

- **Minimum bin frequency**: the smallest number of observations *m* that causes a hexagon to be plotted.

- **Randomize angles of petals**: if checked, the first ray in each region is plotted at a random angle rather than always being vertical. This prevents visual stacking of the hexagons.

References

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.

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

Dupont, W.D. and Plummer Jr., W.D. (2003), "Data Distribution Sunflower Plots", Journal of Statistical Software, 8, 1-5.

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