A Suggestion for Sunflower Plots
Author(s): Mark F. Schilling and Ann E. Watkins
Source: *The American Statistician*, Vol. 48, No. 4 (Nov., 1994), pp. 303-305
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2684839
Accessed: 27/01/2010 02:38

REFERENCES

Bock, R. D. (1979), "Univariate and Multivariate Analysis of Variance of Time-Structures Data," in *Longitudinal Research in the Study of Behavior and Development*, eds. Nesselroade, J. R., and Baltes, P. B., New York: Academic Press, pp. 200–204.

Campbell, R. T., Mutran, E., and Parker, R. N. (1986), "Longitudinal Design and Longitudinal Analysis," *Research on Aging*, 8, 480–502.

Davis, S. D., (1993), "A Computer Program for Regression of Repeated Measures Using Generalized Estimating Equations," *Computer Methods and Programs in Biomedicine*, 40, 15–31.

Gallant, A. R. (1975), "Seemingly Unrelated Nonlinear Regressions," *Journal of Econometrics*, 3, 35–50.

Green P. (1984), "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation and Some Robust and Resistant Alternatives," (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 46, 149–162.

Karim, M. R., and Zeger, S. L. (1988), "GEE: A SAS Macro for Longitudinal Data Analysis," Technical Report 674, The Johns Hopkins University, Dept. of Biostatistics.

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, New York: Chapman and Hall.

Nesselroade, J. R., and Baltes, P. B. (eds.) (1979), *Longitudinal Research in the Study of Behavior and Development*, New York: Academic Press.

SAS Institute Inc. (1985), *SAS/IML User's Guide* (version 5 ed.), Cary, NC: Author.

Ware, J. H. (1985), "Linear Models for the Analysis of Longitudinal Studies," *The American Statistician*, 39, 95–101.

Zeger, S. L. (1988), "Commentary," *Statistics in Medicine*, 7, 161–168.

Zeger, S. L., and Liang, K.-Y. (1986), "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, 42, 121–130.

A Suggestion for Sunflower Plots

Mark F. Schilling and Ann E. Watkins

Although sunflower plots are highly effective for displaying bivariate data with coincident observations, they possess certain disadvantages involving graphical perception of data. Moreover, sunflower plots for data that arrive "on line" can be updated only by completely redrawing the affected sunflowers. We propose a variation of the traditional sunflower plot that addresses these issues.

KEY WORDS: Bivariate data; scatterplots.

1. INTRODUCTION

Sunflower plots (Cleveland and McGill 1984) are a popular and useful tool for the graphical display of bivariate data when there exist cases in which multiple observations occur at or near the same plotting location. The purpose of sunflowers is to convey effectively the density of the data in this situation. The most common alternative method simply prints at each point where multiple observations occur, the numeral representing the number n_i of observations if $1 < n_i < 10$ and a special symbol if $n_i \geq 10$. An example of such a graph is the Minitab plot shown in Figure 1, which shows the daily concentrations of ozone (in parts per billion) in Stamford, Connecticut and Yonkers, New York during the period from May 1, 1974 to September 30, 1974 (data from Chambers, Cleveland, Kleiner, and Tukey 1983).

A sunflower plot is constructed for a bivariate data set by printing a point at each location where $n_i \geq 1$, and then, if $n_i > 1$, superimposing a symmetrically arranged set of n_i short spokes centered at the given point. Figure 2 shows the sunflower plot for the data displayed in Figure 1.

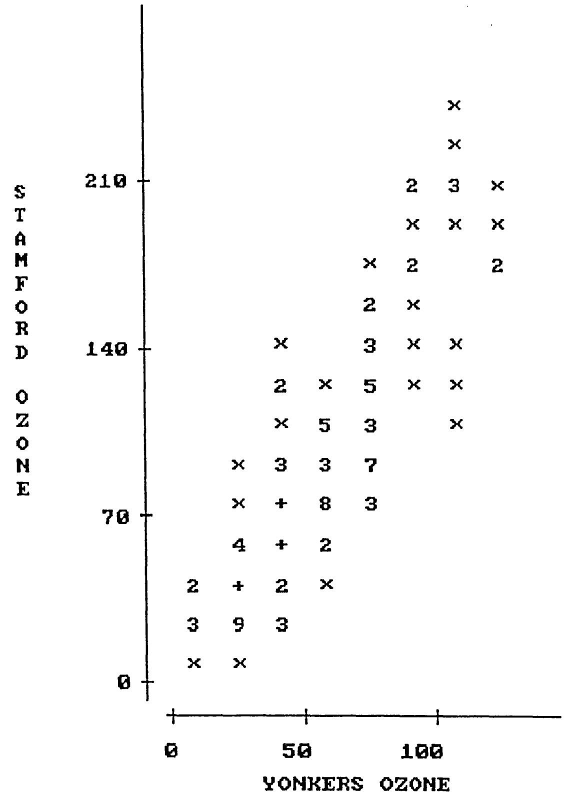

Figure 1. Minitab Scatterplot of Ozone Levels. Data from Chambers, Cleveland, Kleiner, and Tukey (1983). × represents 1 observation and + represents 10 or more.

Although clearly much more effective than Figure 1 in displaying the data, the sunflower plot of Figure 2 possesses certain drawbacks:

(1) Points, which occur when $n_i = 1$, are not sunflowers. It is difficult for the observer to reconcile both points and sunflowers into a unified visual image.

(2) The amount of ink used at each plotting site is not pro-

Mark F. Schilling and Ann E. Watkins are Professors, Department of Mathematics, California State University Northridge, Northridge, CA 91330.

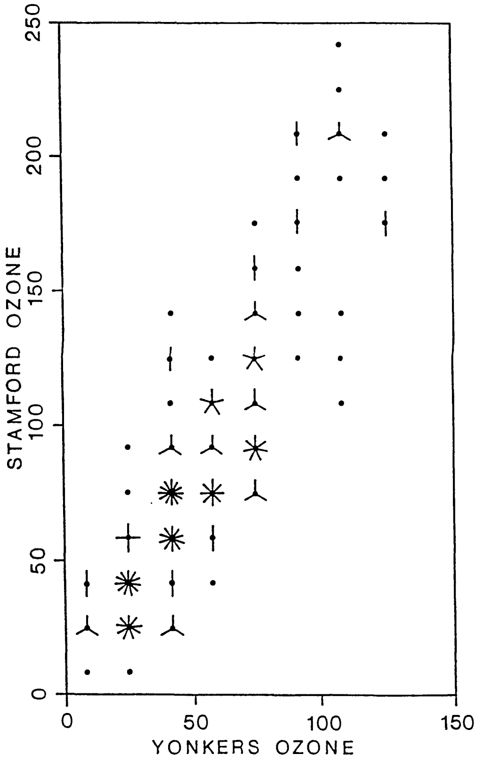

Figure 2. Conventional Sunflower Plot of the Ozone Data of Figure 1. Isolated data points are shown with a dot. Multiple observations are shown with sunflower symbols, in which each spoke represents a data value.

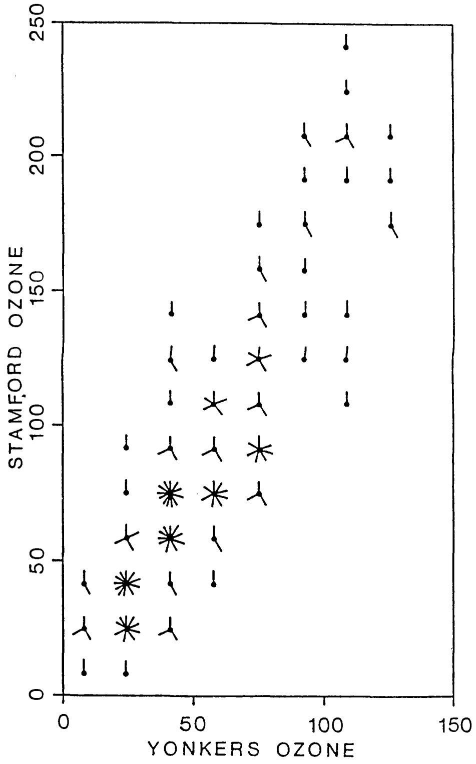

Figure 3. Logarithmic Sunflower Plot of the Ozone Data of Figure 1. Sunflower symbols are used for both single and multiple observations. Spokes possess a logarithmically based angular arrangement and can be generated as data arrive on line.

portional to the number of observations represented there.

(3) For $n_i = 2$, the spokes lie on a single line segment, which may give the subconscious perception that only one observation occurs there, especially if the central point is not large.

(4) The graph cannot be continuously updated as new data become available. Rather, the affected sunflowers must be redrawn, since a sunflower symbol representing m observations generally is not a subimage of a sunflower symbol representing n observations when $n > m$.

2. THE LOGARITHMIC SUNFLOWER PLOT

The first of these four issues can be easily dealt with by using a sunflower with one spoke to represent isolated data points. This approach essentially addresses the second item as well. We propose a scheme that resolves all four of the above issues. The idea arose from thinking about the extreme case of (4): how to make sunflower plots "on line"—that is, when the data arrive sequentially to the plotting program. In this case the objective in forming each sunflower should be to arrange the spokes in such a way that their angular distribution is as even as possible regardless of how many spokes are eventually added to the plot. One of the authors recently developed the following solution for the equivalent problem of sequentially placing points on a circle in such a way that the distribution of points is as even as possible for every possible stopping time of the process.

Let the circle be of unit circumference, for simplicity, and wrap the interval [0, 1] onto the circle, identifying 0 with 1. Let $x_1, x_2, x_3, \ldots$ denote the locations of successively placed points, and let M_n and m_n represent the largest and smallest gaps, respectively, that exist among all adjacent pairs within the first n points. Then among all sequences satisfying certain reasonable properties, the sequence given by $x_k = \log_2(2k-1) \pmod 1$, $k = 1, 2, 3, \ldots$ uniquely minimizes $\limsup_{n\to\infty} nM_n$ and uniquely maximizes $\liminf_{n\to\infty} nm_n$ (Schilling 1992). That is, roughly speaking, this sequence keeps the largest gap as small as possible and the smallest gap as large as possible at all stages of the process. Although these attributes refer to the limiting behavior of the point placing procedure, a good degree of evenness among the gap sizes is achieved by this logarithmic sequence even for very small values of n.

To apply this logarithmic sequence to the construction of sunflower plots, the successive spokes of a sunflower should be drawn so that the kth spoke makes an angle of $2\pi\{\log_2(2k-1) \pmod 1\}$ measured, say, counterclockwise from the vertical. This paradigm will keep the minimum angle between spokes as large as possible—thus enhancing the resolution of distinct observations—in the case when the number of spokes to be drawn is not known in advance.

When a sunflower plot is made by hand from unsorted data, the following simple algorithm can be used to approximate logarithmic sunflowers: Picture a clock face and use only an hour hand to place the spokes as follows: Place spoke #1 at 12:00. For each of spokes #2, #3, #4 and #5, find the time that divides the largest current interval in half (the first interval runs from 12:00 around to 12:00 again) and place the spokes

1 hour earlier, 30 minutes earlier, 15 minutes earlier, and 5 minutes earlier, respectively. For any additional spokes, simply bisect the largest remaining interval each time. The successive spoke times generated are 12:00, 5:00, 8:00, 2:15, 9:55, 6:30, etc., which are remarkably close to the exact logarithmic values. This procedure provides a simple, enjoyable classroom exercise that provides students with hands-on experience in plotting and interpreting bivariate data with coincident observations.

3. EXAMPLE AND DISCUSSION

Figure 3 shows the logarithmic sunflower plot for the data of Figures 1 and 2. The code for creating such a graph is just as simple as that for the standard sunflower plot. While clearly a useful approach for on-line graphing, logarithmic sunflower plotting also addresses issues (1)–(3) for any data set containing multiple observations. Furthermore, there is less geometric regularity in logarithmic sunflower plots than in ordinary sunflower plots, giving a better reflection of the stochastic nature of real data. Finally, the fact that each logarithmic sunflower symbol is a subimage of every logarithmic sunflower symbol with a larger number of spokes tends to synthesize the sunflowers into a more coherent overall plot, allowing the observer to more effectively appreciate the variations in data density.

We feel that the logarithmic approach represents a small but noticeable improvement in graphical perception over the standard method of constructing sunflowers. As it comes at no computational cost, the logarithmic procedure should be considered not just for situations involving the on-line accumulation of data but for all instances in which sunflower plots are appropriate.

REFERENCES

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Boston: Duxbury Press.

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

Schilling, M. F. (1992), "Sequential Partitioning," *The American Mathematical Monthly*, 99, 846–855.

Accent on Teaching Materials

Harry O. Posten, Section Editor

In this section *The American Statistician* publishes announcements and selected reviews of teaching materials of general use to the statistical field. These may include (but will not necessarily be restricted to) curriculum material, collections of teaching examples or case studies, modular instructional material, transparency sets, films, filmstrips, videotapes, probability devices, audiotapes, slides, and data deck sets (with complete documentation).

Authors, producers, or distributors wishing to have such materials announced or reviewed should submit a single, complete copy of the product (three copies of printed material double-spaced) to Section Editor Harry O. Posten, Statistics Department, University of Connecticut, Storrs, CT 06268. A statement of intention that the material will be available to all requesters for a minimum of a two-year period should be provided, along with information on the cost (including postage) and special features of the material. Information on classroom experience may also be included.. All materials submitted must be of general use for teaching purposes in the area of probability and statistics.

Review of the Statistical Education List EdStat-L

EdStat-L is an on-line list concerned with statistical education. Access is free via the Internet. The list originates in the Department of Statistics at North Carolina State University and is operated by Tim Arnold, Director of Instructional Computing in the department.

This review will be of interest mainly to those teaching a service (or first) course in statistics. Suppose that as an inducement to get you to perform this public service, you are offered a chance to attend a national conference on teaching statistics. Suppose the conference even includes significant representation from overseas. There would, no doubt, be many formal sessions and addresses you could attend, but you find that one of the best resources is the convention center cafeteria. The location of the center leaves participants with no other choice for lunch, so each day you walk in, fill a tray, and assign yourself to a table at random. The discussion there is usually lively. As anywhere, some around the table have an opinion on everything, while others rarely speak at all. Some of the discussion seems pretty silly, while some of it is interesting and educational. When the conference is over, you realize you will really miss those lunches. Well, those lunches still continue, although you will have to provide your own food now. On the other hand, you need not have attended the original conference in order to take part.

This parable is the best way I could think of to introduce the email list EdStat-L. Here is how EdStat-L works. Send a message to the list, and that message goes to all the list's subscribers. For example, while preparing this review, I used the list to track down the source of a famous quote from George Box. Within a day or two, I got half a dozen replies. Some were sent to me personally; others were sent to the entire list. This means that if you subscribe to the list, you can benefit from answers given to other people's questions.

According to Tim Arnold, creator of EdStat-L, "the purpose of the forum is to provide a vehicle for comments, techniques and philosophies of teaching statistics. The primary focus is on college-level statistics education, both undergraduate and graduate studies. The forum attempts to bring together every teacher, student, researcher and specialist interested in improving statistics instruction."

EdStat-L started on October 23, 1991 and grew to over 300 members in a matter of days. The forum now has almost 600 email subscribers. The list is also available through the UseNet newsgroup "sci.stat.edu"; approximately 600 to 1,000 people participate in the forum that way. The topics discussed vary tremendously. There have been debates about who should teach statistics and how much probability there should be in a first statistics course. There have been discussions of the purpose of the introductory (and introductory business) statistics course, the role of an experimental