



Opening the Box of a Boxplot

Yoav Benjamini

The American Statistician, Vol. 42, No. 4 (Nov., 1988), 257-262.

Stable URL:

<http://links.jstor.org/sici?sici=0003-1305%28198811%2942%3A4%3C257%3AOTBOAB%3E2.0.CO%3B2-%23>

The American Statistician is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/astata.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

TEACHER'S CORNER

In this department *The American Statistician* publishes articles, reviews, and notes of interest to teachers of the first mathematical statistics course and of applied statistics courses. The department includes the Accent on Teaching Materials section; suitable contents for the section are described

under the section heading. Articles and notes for the department, but not intended specifically for the section, should be useful to a substantial number of teachers of the indicated types of courses or should have the potential for fundamentally affecting the way in which a course is taught.

Opening the Box of a Boxplot

YOAV BENJAMINI*

Variations of the boxplot are suggested, in which the sides of the box are used to convey information about the density of the values in a batch. The ease of computation by hand of the original boxplot had to be sacrificed, as the variations are computer-intensive. Still, the plots were implemented on a desktop personal computer (Apple Macintosh), in a way designed to keep their ease of computation by computer. The result is a dynamic display of densities and summaries.

KEY WORDS: Density estimation; Dynamic graphics.

1. THE BOXPLOT

The boxplot is a simple yet powerful tool for displaying a single batch of data. As a flexible exploratory-data-analysis tool it is used to display data; to study symmetry, "longtailedness," and distributional assumptions; to compare parallel batches of data; and to supplement more complex displays with univariate information. Originated by Tukey (1977), who called it a "schematic plot," it was preceded by the box-and-whiskers plot also described in the same book. It is now customary to refer to this second version of the plot as a boxplot.

The boxplot displays a rectangle oriented with the axes of a coordinate system in which the vertical axis has the scale of the batch of data. Its top and bottom are drawn at the upper and lower quartiles of the batch. This box is cut by a horizontal line at the median. A step is defined as 1.5 times the interquartile range, and a vertical line is extended from the middle of the top of the box to the largest observation within a step from the top. A similarly defined line extends from the bottom of the box to the smallest observation within a step from the bottom. Observations that are farther from the box than these lines are individually displayed; those that are more than two steps away from the box are graphically emphasized. The definition of the quartiles might vary, and sometimes multipliers other than 1.5 are used (or are recommended for use) in the definition of

a step. Hoaglin, Iglewicz, and Tukey (1986) explored in depth this last issue, in terms of its implications on the number of individually displayed points. Frigge, Hoaglin, and Iglewicz (1987) addressed the definition of quartiles.

What are the properties of the boxplot that make it so useful?

1. Five summaries of the data are graphically presented in a way that makes the information about the location, spread, skewness, and longtailedness of the batch available with a quick glance. To be more specific, location is displayed by the cut line at the median (as well as by the middle of the box), spread by the length of the box (as well as by the distance between the ends of the whiskers and the range), skewness by the deviation of the median line from the center of the box relative to the length of the box (as well as by the length of the upper whisker relative to the length of the lower one, and by the number of individual observations displayed on each side), and longtailedness by the distance between the ends of the whiskers relative to the length of the box (as well as by the number of observations specially marked).

2. The boxplot displays detailed information about the observations at the tails. If there is an interest in the value of an observation, it is usually one in the tail. In fact the boxplot is a compromise between a detailed description of a batch of data such as a line plot (jittered or not) or a stem-and-leaf plot and a condensed display of summaries only (see Sec. 2).

3. The distributions of many batches of data can be easily compared by displaying their boxplots side by side.

4. The boxplot is easy to compute. It was designed as a back-of-the-envelope graphical method, and it easily lends itself for implementation on computer devices with no special graphical capabilities; a line printer is enough.

5. The meaning of a boxplot can be easily explained to users of statistics. Plotting one boxplot by hand with a scientist is in most cases all of the explanation needed. Usually users later draw boxplots themselves and incorporate them into their publications. The only question I sometimes get is: How wide should the box be?

2. ALTERNATIVES AND VARIATIONS

There are other ways to display a single batch of data: stem-and-leaf display, histogram, and density trace, to name a few. Plotting the points along a line—a line plot—is the

*Yoav Benjamini is Lecturer, Department of Statistics, Sackler School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel. The research for this article was partially done while the author was visiting the Department of Statistics at the Wharton School of the University of Pennsylvania. The author thanks Yair Benjamini for his help in programming and Allan R. Wilks for helpful comments.

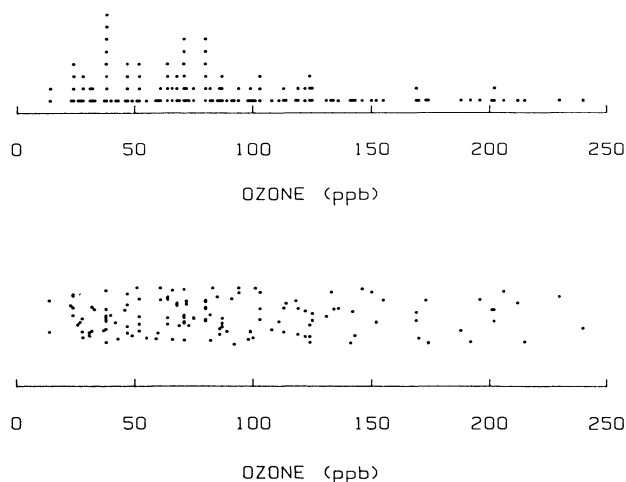


Figure 1. Stacked and Jittered Line Plot (from Chambers et al. 1983). Reprinted by permission of PWS-Kent Publishing Company.

simplest. To solve the problem of the cluttering of points, the points can be stacked or randomly jittered along a second dimension. Examples can be seen in Figure 1. The common feature of these plots is that they show the distribution of the batch in greater detail but do not include any graphical description of summaries. For more on these methods see Chambers, Cleveland, Kleiner, and Tukey (1983).

A variation of the boxplot is the notched boxplot of McGill, Larsen, and Tukey (1978). The regular boxplot is supplemented by an approximate confidence interval for the median of the batch, shown as a pair of wedges taken out of the sides of the box. These confidence intervals are constructed in such a way that when two notches of different boxplots do not overlap their medians are significantly dif-

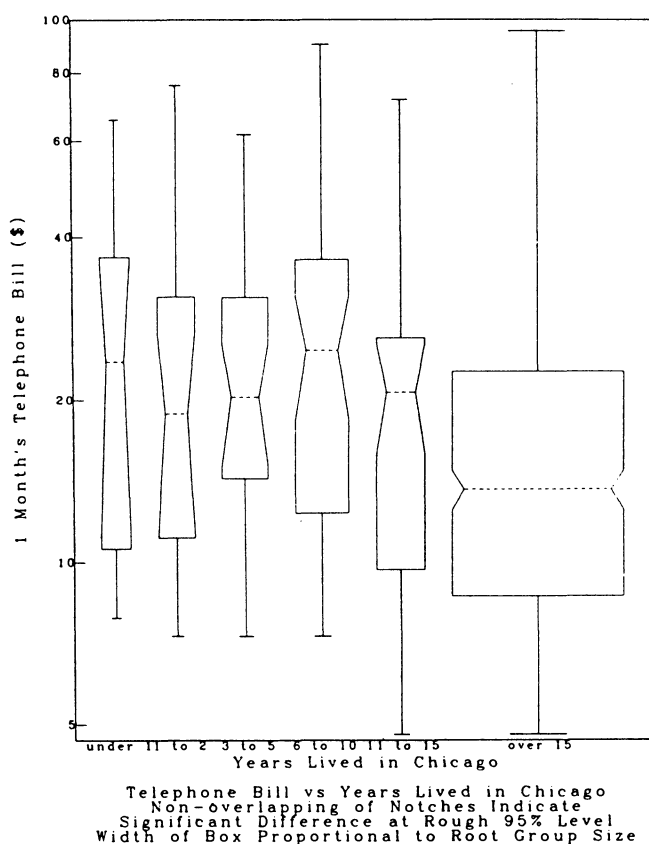


Figure 2. Notched Boxplots (from McGill et al. 1978).

ferent. A second variation makes the width of the box proportional to the square root of the batch size. Figure 2 gives an example of both (though in the earlier form of the box-and-whiskers plot with lines extending to the extremes). Since the formula for the confidence interval is a constant times the interquartile range divided by the square root of the batch size, the latter can be perceived from the length of the wedges relative to the length of the box.

Tufte (1983) recommended stripping the boxplot down to its essentials. He finds no use for the box, as only its upper and lower positions are needed. Thus, to increase the data-ink ratio, he suggests a plot as in Figure 3.

3. OPENING THE BOX

Looking at Figure 3 gives the strange impression of seeing no data where the data are actually mostly concentrated. This uneasy feeling does not arise in the original boxplot because the box gives the visual impression of more concentration of data at the center. Realizing that we do interpret intuitively the width of the box as a sign of concentration, this approach can be carried one step further. I suggest using the sides of the box to portray density information about the data. Next I present two implementations of this idea.

3.1 The Histplot

A simple way of incorporating information about the density into the boxplot is to estimate the density at the median and at the two quartiles. Then draw the top, bottom, and median line of the box with width proportional to the estimated density, and connect them by straight lines as in a frequency polygon. Farther outside, the details of the boxplot are unchanged. The result may be called a *histplot*, as it is a hybrid of a five-bin histogram and a boxplot. The resulting shape of the central part is that of two equilateral trapezoids sharing a common base at the median. Because only the relative widths matter, essentially two new summaries have been added to the boxplot display. If the estimated density of the three quantiles is defined in terms of inverse of distances between octiles (i.e., bin edges at the .125, .375, .625, and .875 quantiles), this plot can practically be used for samples of sizes as small as 9. In general any other method can be used for estimating the density, such as kernel estimation or k nearest neighbors. The computational aspects are further discussed in Section 5.

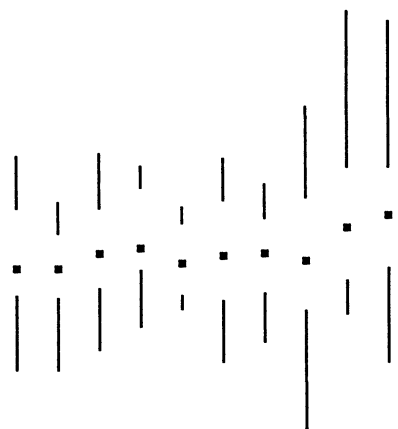


Figure 3. Tufte's (1983) Version of the Boxplot. Reprinted by permission.

Three samples are presented using histplots in Figure 4. Each is a random sample of size 40 from a chi-squared distribution with 1, 3, and 4 degrees of freedom (shifted and rescaled). It is interesting to note how the additional information gotten by "opening the box" enables us to capture from these small samples the differences in skewness between these chi-squared distributions.

If a confidence interval about the median is needed, the width of the box cannot be used for displaying it as in the notched boxplot. In Figure 7c (see Sec. 4) these are shown by the gray bar extending from the median. As a bonus we find them more easily comparable across separated histplots than the notches of McGill et al. (1978).

3.2 The Vaseplot

A *vaseplot* is a boxplot where the width of the box at each point is proportional to the estimated density there. As a result the box is replaced by a vase-like shape. (The more artistically inclined can visualize the vase, the stem, and some outlying flowers in the display.) An example of the vaseplots of the three samples described in Figure 4 can be seen in Figure 5.

The shape of the central part is highly dependent on the method of estimation of the density. In particular, the most important parameter is the one governing the smoothness of the estimated density. In the nearest-neighbors method, where one uses the length of the smallest interval to contain k observations, it is the number of observations k . In kernel estimation it is the width of the kernel window h . Since the latter is the method used for generating the vaseplots presented here, the expression for the estimated density at x is written explicitly as

$$f(x) = \sum_i W((x_i - x)/h)/hn,$$

where W is a symmetric window function (e.g., cosine, Gaussian, bisquare, or simply boxcart) and h is the window width. See Chambers et al. (1983) for details and for further references to the vast literature on density estimation. In the

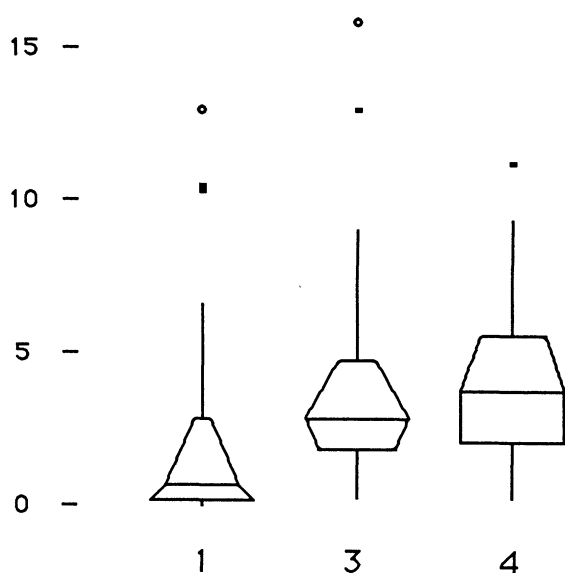


Figure 4. Histplots of Three Samples of Size 40 From Linearly Transformed Chi-Squared Distributions With 1, 3, and 4 Degrees of Freedom.

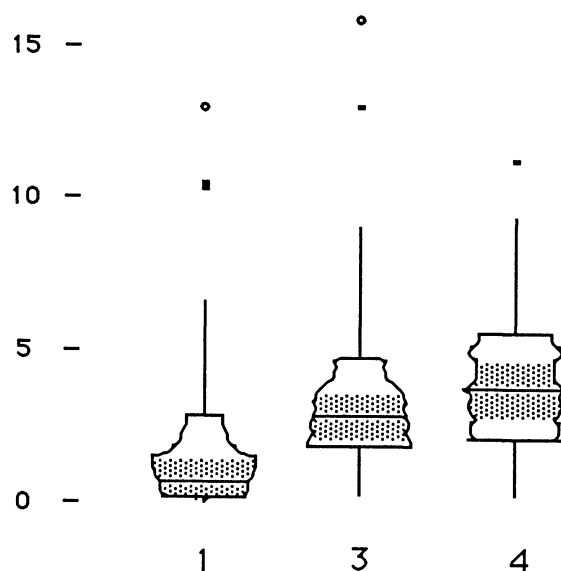


Figure 5. Vaseplots of Three Samples of Size 40 From Linearly Transformed Chi-Squared Distributions With 1, 3, and 4 Degrees of Freedom. The gray bars are confidence intervals for the medians.

case of a boxcart window,

$$W(u) = 1 \quad \text{for } |u| \leq .5 \\ = 0 \quad \text{for } |u| > .5.$$

One can easily see that for h larger than the range of the data a constant density is assured, so the conventional boxplot is a limiting form of the vaseplot. On the other hand, for a very small h the density is bound to show a spike for each observation. If h is exactly at the value that is also the resolution of the display, the resulting vaseplot takes the form of a line plot where the multiplicity of the points is coded by the width of the line at the point. Another way of describing this last plot is as a histogram with very narrow bin width. Thus the vaseplot can be thought of as a plot with a continuum of possible displays for the center of the distribution, ranging from the least informative boxplot to the possibly most detailed histogram. This range of possibilities is explored on the same data set in Figure 6.

The question of which h should be used to get the best (in some sense) description of the density is a complex one. Though many iterative methods exist, I like the "rabbits' ears hunting" approach: Start with a large window width h , and decrease it to get a slowly varying yet smooth density trace. A density trace that looks as if rabbits are hiding behind the hills (and only their ears are appearing) is a sign of going too far; return to the last h before the appearance

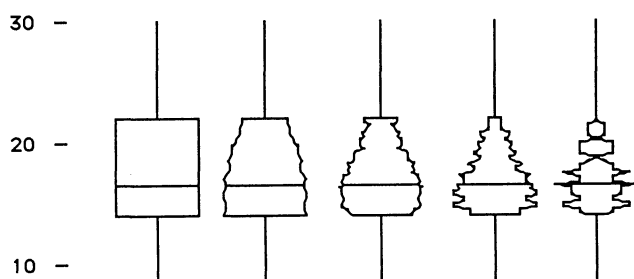


Figure 6. Five Vaseplots of the Data From Figure 5. The window width decreases from left to right. The cosine window (kernel) is used.

of the roughness. It is important to emphasize that there need not be one single best representation, and different window widths can be used for different purposes.

I found that the most appealing way for displaying the confidence intervals for the medians was shading the appropriate part of the vase. When needed, the shading extends beyond the vase. I also found that for the more detailed versions, where the sides of the vase tracing the density become quite rough, shading the inside of the vase enhances the ease of interpretation. If in such cases confidence intervals are needed, they are shaded darker. I did not choose to always shade the inside of the vase because this reduces the visibility of the median line.

I should emphasize why I chose to modify only the width of the box and not display the density over the whole range of the data. There are three reasons:

1. It is essential to keep an easily accesible graphical display of the summaries of the batch.

2. By tracing the density all of the way to the tails one loses the information about individual observations in the tails.

3. Large sample sizes are needed for estimating the usually low density at the tails. I restrict the density description to the region where it is usually higher, and thus can be better estimated.

4. A PROTOTYPE OF DYNAMIC COMPUTER IMPLEMENTATION

Unfortunately, one important advantage of the boxplot has not been retained by the vaseplot: It should by now be clear that the vaseplot is not easy to compute. The purpose of this section is to show that although we cannot keep the ease of computation by hand, we can still easily draw vaseplots with a computer. The discussion in Section 3, about the flexibility of the vaseplot, suggests that the appropriate environment for creating vaseplots should be interactive.

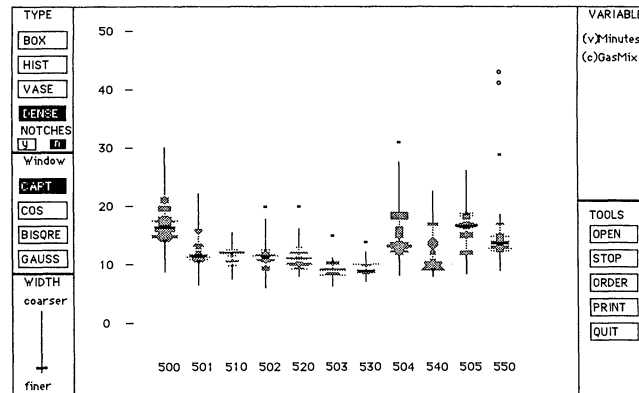
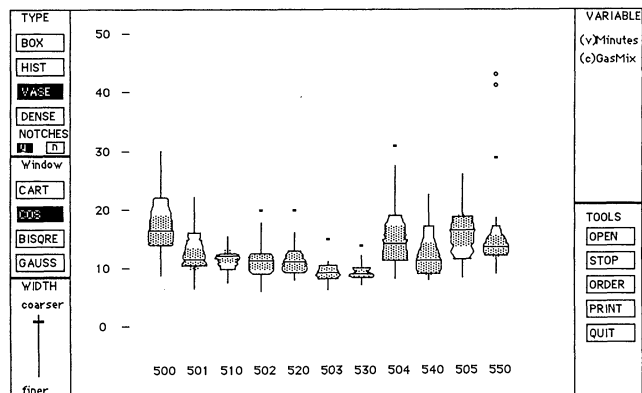
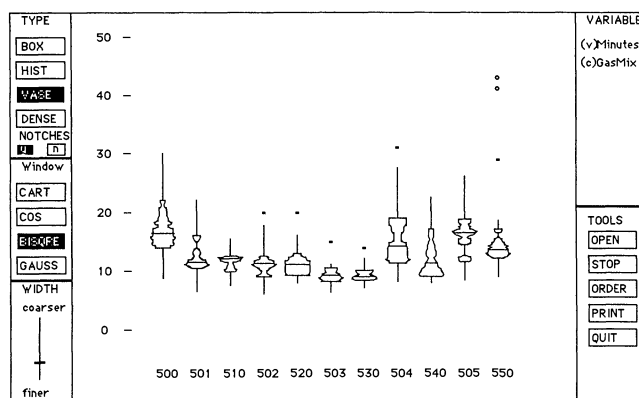
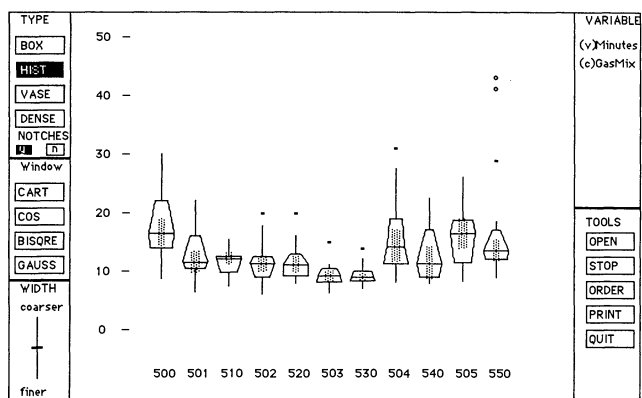
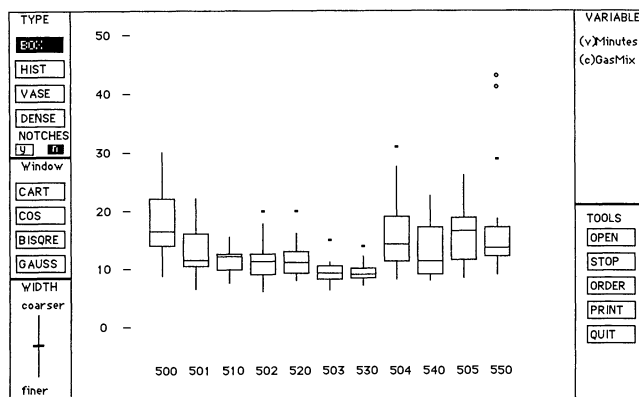
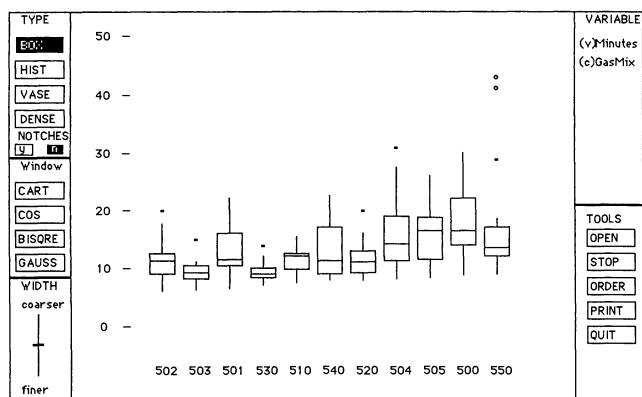


Figure 7. Time (in minutes) Until Start of Convulsions in Rats Breathing Pressurized Oxygen Mixed With Either Pressurized Nitrogen or Helium.

The choice whether to use boxplots, histplots, or vaseplots should be made in view of the consequences of the choice. Even when the desired display is chosen there are further choices to be made for each display: Are confidence intervals needed? Should the vase be shaded? What type of window is preferable for the density estimation? How smoothly should the density be traced?

The last question suggests that to make a good choice the display should be highly interactive or dynamic [in the sense of Becker, Cleveland, and Wilks (1987)], allowing instant feedback as window width is changed. The rabbits'-ears-hunting strategy discussed earlier for density estimation is critically dependent on having such a capability. Additional operations that are desirable for displaying batches of data are the ability to show multiple plots of the same variable at different levels of a second one, merging and separating groups, choosing the order of the displayed batches, and transforming the variables. From the arsenal of the dynamic methods discussed by Becker et al. (1987), it is desirable to have labeling (where the information about the outside observations could be displayed upon request) and deletion (where an observation pointed at is deleted). One might also want to merely suppress the drawing of a point, even though it is still included in the computation process, so that the rest of the display is not overly condensed. Tukey (1977, p. 206) gave nice examples of this approach.

The presentation of more than one vaseplot at a time raises some complications. The first one is whether the densities should be plotted on the same scale across the different vases. I chose to rescale each vaseplot separately so that the maximum density for each batch is coded to the same width. If densities are plotted on a common scale, one batch with small spread will suppress the details of the vases in the batches with larger spread. The second issue is whether to use one window width for all jointly displayed vaseplots or to choose it separately for each one. Theoretical results (e.g., Tapia and Thompson 1978) show that, for a random sample, the optimal window width should decrease with increasing sample size at the rate of the fifth root of the sample size. Furthermore, if two distributions differ only by scale, the optimal window widths are proportional to the scales. In general the vaseplots are clearly not restricted to displaying random samples, but the preceding points could serve as guidelines. Therefore, I decided to use a single handle for choosing the smoothing parameter simultaneously for all of the plots (see the "width" box in Fig. 7). It is then multiplied for each batch by its interquartile range and divided by the fifth root of the number of observations of the batch to automatically determine the individual window width.

These ideas were implemented on the Apple Macintosh desktop computer, making use of its point-and-click mouse interaction and its high-resolution display. The program was written in Macintosh Pascal, with the help of Yair Benjamini. It is not meant to be a stand-alone data-analysis program, but a prototype for a program to be incorporated into a good interactive environment for data analysis. (Some of these already exist for the Macintosh.) Consequently, we were not too concerned with speed. A method for fast density estimation while changing the window width, such as

the one suggested by Scott (1987), might be useful for a practical implementation of the vaseplot. In the next section I illustrate the use of the program and the plots for the analysis of some data.

5. AN EXAMPLE

To give a feeling of the way the program works, Figure 7 comprises six still shots of screens taken (by clicking on the "print" menu bar in the tool box on the bottom right) while analyzing some data. The observations are taken from a study by Bitterman, Laor, and Melamed (1987). Each observation is the time until start of convulsions in a rat breathing a pressurized mixture of gases and the type of mixture it was breathing. Five atmospheres of oxygen were administered to rats in one group. Five atmospheres of oxygen mixed with one to five atmospheres of either nitrogen or helium were administered to the other groups. There were 18–28 rats exposed to each of the 11 mixtures.

Reading in the data involves clicking on the menu bar "open" in the bottom right tool box, and then defining the file with the data. The file contains two variables for each observation: the time to convulsions and an identification number for the type of mixture. This identification number has three digits: the left digit is the number of atmospheres of oxygen, the center digit is that of helium, and the right digit is that of nitrogen. The two variables are then listed in the upper right data window. One variable is chosen to be displayed (v), and another is chosen as a grouping variable (c).

First, we want to get the simplest display of these batches. Clicking on the "box" menu bar, in the type box at the upper left corner, produces the screen in Figure 7a. The batches do not seem to share the same distribution, and we would like to further explore these differences.

It is generally known that the higher the pressure the shorter the time to convulsions is. We would like to check if this can explain the differences in locations among the batches. To facilitate the visual comparison we would like to arrange the batches in order of increasing overall pressure of the mixture. (The overall pressure is the sum of the three digits.) Clicking inside the "order" menu bar (down right) opens small windows underneath each group. Clicking according to the desired order inside the windows does the job.

The pattern emerging in Figure 7b is clear and interesting. But before elaborating on its meaning we want to be more confident that the observed differences in the locations of the medians are not merely due to the random fluctuations. So we introduce the confidence intervals for the medians by clicking inside the notches "y" bar (on the left). We would also like to have a more detailed look at the distribution of the observations in the centers. So we choose to display histplots instead of boxplots, and we get them by clicking on "hist" (on the upper left). Instantaneously, the screen appears as in Figure 7c.

It is evident now that the time to convulsions does not simply decrease as the total pressure increases. The relationship is nonmonotone as well as nonlinear. Time to convulsions first decreases sharply, then more slowly, and finally rises again—in an almost quadratic behavior over this range.

We also notice that a higher median time to convulsions is associated with a larger spread within the group. Skewness is also more evident in these groups. It seems that there is some difference between each pair of mixtures having the same pressure; to display the batches in greater detail we would like to draw their vaseplots.

Click inside "vase" in the upper left side, and the vaseplots are drawn using the default type of window—the boxcart. It can be changed to a bisquare window, for example, by clicking inside the (central left) "bisqre" bar, and the display changes to reflect the differently estimated density. Changing the window width is done by changing the location of the handle on the scale in the (lower left) width box via a click with the mouse. This will dynamically change the displayed vaseplots. Figures 7d and 7e show the results of two settings. We now notice a difference in shape between mixtures containing nitrogen and helium: The latter tend to have stronger and more consistent skewness evident, not only in the tails but also throughout the center.

Finally, if we want to study the observations in detail while giving up some of the global view, we choose a very narrow window width and choose "dense," which fills up the vase. The display we get (Fig. 7f) can clearly function in lieu of a table containing the values and still have most of the advantages of the boxplots in Figure 7c.

At this point we might want to repeat the analysis using the (possibly inversely) transformed data. When a program of this type is implemented in a data-analysis environment, this should not be a problem. We still might like to view

dynamically the resulting plots from transforming the observations—for example, as we choose various powers from the power family of transformations.

[Received July 1987. Revised January 1988.]

REFERENCES

- Becker, R. A., Cleveland, W. S., and Wilks, A. R. (1987), "Dynamic Graphics for Data Analysis," *Statistical Science*, 2, 355–382.
- Bitterman, N., Laor, A., and Melamed Y. (1987), "CNS Oxygen Toxicity in Oxygen-Inert Gas Mixtures," *Undersea Biomedical Research*, 14, 477–483.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth.
- Frigge, M., Hoaglin, D. C., and Iglewicz, B. (1987), "Some Implementations of the Boxplot," in *Computer Science and Statistics: Proceedings of the 19th Symposium on the Interface*, Alexandria, VA: American Statistical Association, pp. 296–300.
- Hoaglin, D. C., Iglewicz, B., and Tukey, J. W. (1986), "Performance of Some Resistant Rules for Outlier Labeling," *Journal of the American Statistical Association*, 81, 991–999.
- McGill, R., Larsen, W. A., and Tukey, J. W. (1978), "Variations of Box Plots," *The American Statistician*, 32, 12–16.
- Scott, D. W. (1987), "Software for Univariate and Bivariate Average Shifted Histogram," Technical Report 311-87-1, Rice University, Dept. of Mathematical Sciences.
- Tapia, R. A., and Thompson, J. R. (1978), *Nonparametric Probability Density Estimation*, Baltimore: Johns Hopkins University Press.
- Tufte, E. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.
- Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.