Violin Plots: A Box Plot-Density Trace Synergism

Jerry L. HINTZE and Ray D. NELSON

Many modifications build on Tukey's original box plot. A proposed further adaptation, the violin plot, pools the best statistical features of alternative graphical representations of batches of data. It adds the information available from local density estimates to the basic summary statistics inherent in box plots. This marriage of summary statistics and density shape into a single plot provides a useful tool for data analysis and exploration.

KEY WORDS: Density estimation; Exploratory data analysis; Graphical techniques.

1. INTRODUCTION

Many different statistics and graphs summarize the characteristics of single batches of data. Descriptive statistics give information about location, scale, symmetry, and tail thickness. Other statistics and graphs investigate extreme observations or study the distribution of data values. Diagrams such as stem-leaf plots, dot plots, box plots, histograms, density traces, and probability plots give information about the distribution of values assumed by all observations.

The violin plot, introduced in this article, synergistically combines the box plot and the density trace (or smoothed histogram) into a single display that reveals structure found within the data. The introduction of this new graphical tool begins with a quick overview of the combination of the box plot and density trace into the violin plot. Then, three illustrations and examples show the advantages and challenges of violin plots in data summarization and exploration.

2. COMPONENT PARTS OF VIOLIN PLOTS

The violin plot, as depicted in Figure 1 and implemented in NCSS (1997) statistical software, combines the box plot and density trace into one diagram. The name *violin plot* originated because one of the first analyses that used the envisioned procedure resulted in a graphic with the appearance of a violin. Violin plots add information to the simple structure of the box plot that Tukey (1977) initially conceived. Although these original graphs are easily drawn with pencil and paper, computers ease subsequent modifications, refinements, and computation of box plots as discussed by McGill, Tukey, and Larsen (1978); Velleman and Hoaglin (1981); Chambers, Cleveland, Kleiner, and Tukey (1983); Frigge, Hoaglin, and Iglewicz (1989), and others.

Box plots show four main features about a variable: center, spread, asymmetry, and outliers. As an example, consider the box plot in Figure 1 for the data published by Hamermesh (1994). The ASA Statistical Graphics Section's 1995 Data Analysis Exposition analyzes these data, which report compensation of professors from all academic ranks in the United States. The labels in the diagram identify the principal lines and points which form the main structure of the traditional box plot diagram. As shown, the violin plot includes a box plot with two slight modifications. First, a circle replaces the median line which facilitates quick comparisons when viewing multiple groups. Second, outside points which are traditionally classified as mild and severe outliers, are not identified by individual symbols.

The density trace supplements traditional summary statistics by graphically showing the distributional characteristics of batches of data. One simple density estimator, the histogram, displays the distribution of data values along the real number line. Weaknesses of the histogram caused Tapia and Thompson (1979), Parzen (1979), Silverman (1986), Izenman (1991), and Scott (1992) to propose and summarize numerous alternative density estimators. One of these alternatives is the density trace described in Chambers, Cleveland, Kleiner, and Tukey (1983). Defining the location density $d(x|h)$ at a point x as the fraction of the data values per unit of measurement that fall in an interval centered at x gives

$$d(x|h) = \frac{\sum_{i=1}^{n} \delta_i}{nh}, \tag{1}$$

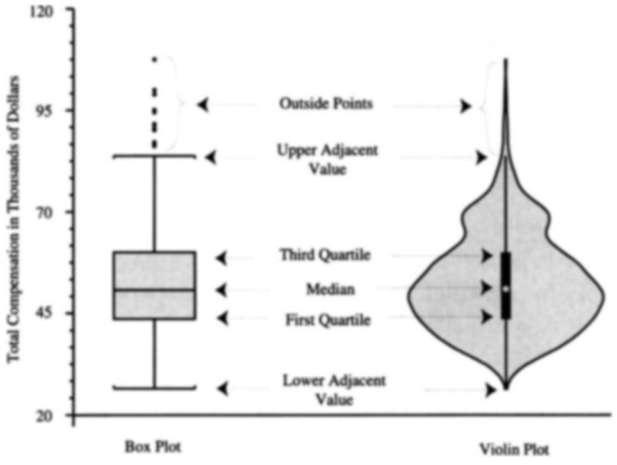

Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

Jerry L. Hintze is President, NCSS, 329 North 1000 East, Kaysville, UT 84037 (E-mail: sales@ncss.com). Ray D. Nelson is Associate Professor of Business Management, Marriott School of Management, Brigham Young University, Provo, UT 84602.

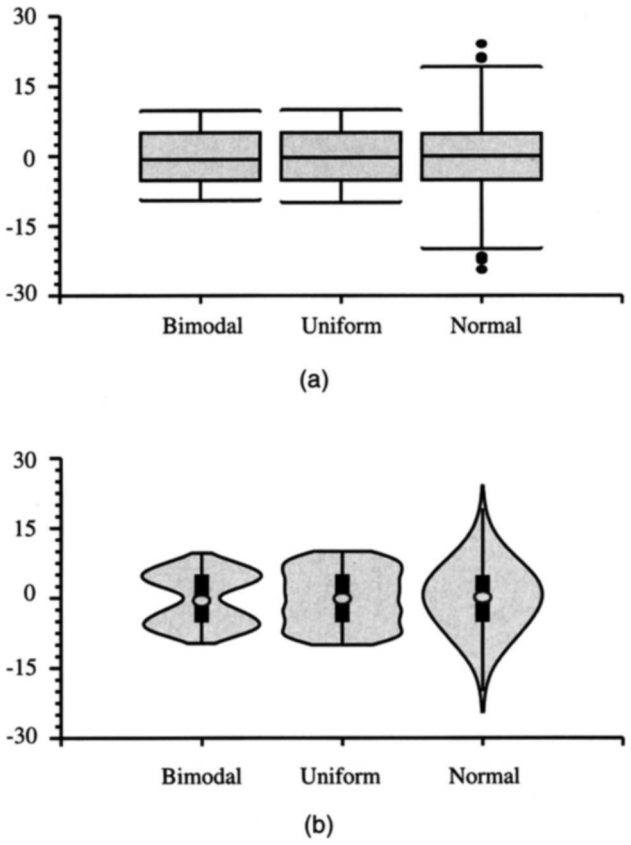

Figure 2. Comparison of Box Plots and Violin Plots fo Known Distributions. (a) Box plots; (b) violin plots.

where n is the sample size, h is the interval width, and δ_i is one when the ith data value is in the interval $[x - h/2, x + h/2]$ and zero otherwise. In order to plot the density trace, first select a value for h and then compute $d(x|h)$ on a dense grid of equally spaced x values. Connect the $d(x|h)$ by lines. The shape of the $d(x|h)$ curve is essentially driven by the interval length, h. It is very smooth for large values of h, and "wiggly" for small values.

Unfortunately, several density traces shown side by side are difficult to compare. Contrasting the distributions of several batches of data, however, is a common task. In order to add information to the box plot and still make comparisons possible, Benjamini (1988) suggested "opening the box" of the box plot. He makes the width of the box proportional to the estimated density. The violin plot builds on the Benjamini proposal by combining the advantages of box plots with density traces.

The violin plot, as shown in Figure 1, combines the box plot with density traces. The density trace is plotted symmetrically to the left and the right of the (vertical) box plot. There is no difference in these density traces other than the direction in which they extend. Adding two density traces gives a symmetric plot which makes it easier to see the magnitude of the density. This hybrid of the density trace and the box plot allows quick and insightful comparison of several distributions.

3. SPECIFICATION OF INTERVAL WIDTH

As with other density estimators, achieving an acceptable density trace requires experience and judgment in determining the appropriate amount of smoothing. As with the selection of the bin width in the histogram, the interval width h, which is usually specified as a percentage of the data range, must be selected. Experience suggests that values near 15% of the data range often give good results. The choice of h, however, must be tempered by the size of the sample. The density trace is subject to the same sample size restrictions and challenges that apply to any density estimator. For small data sets, too small a value for h gives a wiggly density trace that suggest features that are simply artifacts of the individual data points. The oversmoothed density estimate that results from too large h values gives the illusion of knowing the shape of the distribution, while in reality the data set is too small for any conclusions. As a rule of thumb based on practice, the density trace tends to do a reasonable job with samples of at least 30 observations. Even with sample sizes of several hundred, however, choosing too large a value for h causes the density trace to oversmooth the data. In general, values of h greater than 40% of the range usually result in oversmoothed densities, while values less than 10 percent of the range result in undersmoothed densities. Hence, percentages between 10 and 40 percent are recommended.

4. ILLUSTRATIONS AND APPLICATIONS

With the addition of the density trace to the box plot, violin plots provide a better indication of the shape of the distribution. This includes showing the existence of clusters in data. The density trace highlights the peaks, valleys, and bumps in the distribution. Three applications and examples of violin plots illustrate these advantages. The first example demonstrates the ability of violin plots to distinguish among the shapes of known distributions. The second highlights

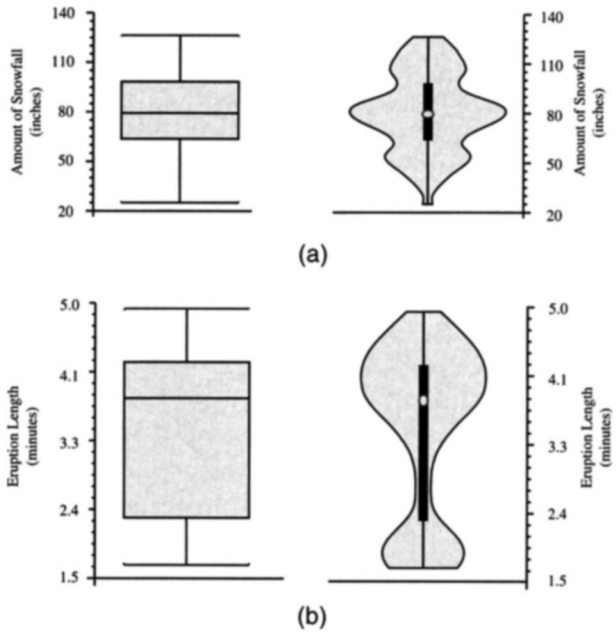

Figure 3. Additional Information in Violin Plots. Two examples from the density estimation literature: (a) annual snowfall for Buffalo, NY, 1910—1972; (b) Old Faithful eruption length.

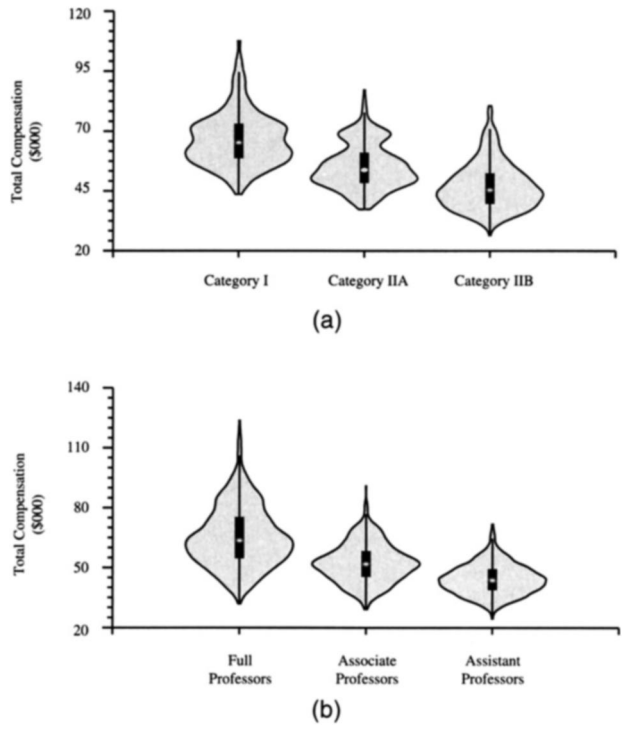

Figure 4. Exploring Data with Violin Plots. Total compensation data from ASA analysis competition. (a) Compensation of all professors by university classification; (b) total compensation by academic rank.

the ability to detect bumps or clusters of data. The third shows their potential in exploring for structure and pattern in the academic compensation data used previously in the illustration of the components of violin plots. The values for the interval widths h are chosen using personal judgment from values from within the recommended 10 to 40 percent interval. These examples establish the potential of violin plots in data analysis and exploration.

4.1 Comparison of Known Distributions

Consider first the ability to detect general shapes for distributions of data. Figure 2 depicts box plots and violin plots for random samples of 10,000 simulated observations drawn from three different known distributions. The three distributions share identical location and scale characteristics as measured by the median and interquartile range. The first is a bimodal distribution with modes at -5 and 5 and range between -10 and 10. The second is a uniform distribution on the interval $[-10, 10]$. The third is a normal distribution $N(0, 54.95)$. The box plots in Figure 2(a) reflect the fact that all three have the same median and interquartile range.

As expected, the density trace accurately reveals the shape of the distribution from which the random samples are drawn. The violin plot for the bimodal distribution clearly shows the twin peaks of the known distribution. Unfortunately, box plots cannot differentiate between the shapes of the bimodal and uniform distributions. The box plots do, however, show that the normal distribution differs from the others as it does have a larger range.

These plots seem to indicate that since the mass of the bimodal plot is less than the normal plot, the bimodal plot is based on fewer observations. This is a weakness of this implementation of the violin plot, which adjusts the density traces so that their maximum heights are equal. This allows a direct comparison of the shapes, but removes the visual impact of sample size. A variation of this implementation would keep a uniform scaling of the density traces.

4.2 Density Estimation Examples

Clusters of data appear as bumps in density estimators. Box plots often do not alert analysts to their existence. Two examples from the density estimation literature clearly illustrate this ability. First, Parzen (1979) and Scott (1992) used annual snowfall data for Buffalo, New York for 1910–1972 to show the value of nonparametric density estimation. The violin plot in Figure 3(a) illustrates the additional insights available through density estimators that the basic box plot does not reveal. The second example in Figure 3(b), which uses data previously considered by Silverman (1986) and Scott (1992), shows the bimodal nature of Old Faithful eruption lengths. Once again, the violin plot clearly adds significant insight about the distribution of the process generating the data.

4.3 Academic Compensation

The information that violin plots add to box plots increases the potential of these tools when used in data exploration. As an example of the value of violin plots, consider the diagrams in Figure 4 for data published by Hamermesh (1994). The graphics in Figure 4(a) summarize total compensation of all professors for three different classifications of colleges and universities. The first category includes institutions with a significant level of doctoral-level education. The second encompasses institutions with diverse post-baccalaureate programs but which do not have a significant level of doctoral programs. The colleges and universities in the third category focus their primary activity on undergraduate baccalaureate-level education. The bumps in the doctoral level and post undergraduate violin plots suggest that some universities in each of these categories might have compensation characteristics which distinguish them from other members of the general group.

Figure 4(b) shows the distribution of total compensation for institutions in all three categories by academic rank. All three of the categories appear to be somewhat positively skewed with the skewness increasing with the academic rank. Comparison of the medians gives the expected increase in compensation with the higher academic rank. An interesting bulge grows in the upper tail of the distributions as the academic rank increases. In an exploratory analysis, the violin plots point to the next question which might investigate the characteristics of the institutions in these clusters.

5. SUMMARY AND CONCLUSIONS

Individually, box plots provide succinct summaries of data. By themselves, density traces reveal important information about the distribution of data. The synergistic com-

bination of the box plot and the density trace allows much of the information from each to be displayed in one plot. This single plot structure makes comparisons of distributional factors of several variables much easier. Three different illustrations show that violin plots retain much of the information of box plots and add information about the shape of the distribution not obvious in box plots. Their ability to detect clusters or bumps within a distribution is especially valuable.

[Received February 1997. Revised November 1997.]

REFERENCES

Benjamini, Y. (1988), "Opening the Box of the Box Plot," *The American Statistician*, 42, 257–262.

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. (1989), "Some Implementations of the Box Plot," *The American Statistician*, 43, 50–54.

Hamermesh, D. (1994), "Plus Ça Change: The Annual Report on the Economic Status of the Profession, 1993–1994," *Academe*, 5–89.

Hintze, J. (1997), *User's Guide, NCSS 97, Statistical System for Windows*, Kaysville, UT: Number Cruncher Statistical Systems (http://www.ncss.com).

Izenman, A. J. (1991), "Recent Developments in Nonparametric Density Estimation," *Journal of the American Statistical Association*, 86, 205–224.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978), "Variations of Box Plots," *The American Statistician*, 32, 12–16.

Parzen, E. (1979), "Nonparametric Statistical Data Modeling," *Journal of American Statistical Association*, 74, 105–131.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: Wiley.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.

Tapia, R. A., and Thompson, J. R. (1978), *Nonparametric Probability Density Estimation*, Baltimore, MD: Johns Hopkins University Press.

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

Velleman, P. F., and Hoaglin, D. C. (1981), *Applications, Basics and Computing of Exploratory Data Analysis*, Boston: Duxbury Press.