A Versatile One-Dimensional Distribution Plot:
The BLiP Plot

J. Jack Lee and Z. Nora Tu

A versatile graphical tool, the BLiP plot, was developed for displaying one-dimensional data. The basic building blocks are boxes, lines, and points. Like many standard one-dimensional distribution plots, the BLiP plot is capable of displaying individual data values in points or lines and grouped information in lines or boxes. In addition, the BLiP plot includes many new features such as variable-width plots and several choices of point patterns. The main advantage of the BLiP plot is that it provides users with basic graphical elements in a friendly and flexible environment so that users can, according to their needs, construct anything from a simple, standard plot to a complex, customized plot to best present their data.

KEY WORDS: Boxplot; Histogram; One-dimensional scatterplot; Summary statistics; Variable-width plots.

1. INTRODUCTION

Standard tools for displaying a sample distribution include stem-and-leaf plots, histograms, frequency polygons, boxplots, one-dimensional (1-D) scatterplots, and variations of these graphs. There are two basic approaches to using these graphical tools for presenting information contained in the data. One approach focuses on the value of each individual observation. The other approach emphasizes grouped information or summary statistics. For example, 1-D scatterplots are more appropriate for showing the locations of individual data points, while histograms and frequency polygons are suitable for revealing the distribution of grouped data. In contrast, stem-and-leaf plots and boxplots have a flavor of both. Stem-and-leaf plots show the value of each data point in the "leaf," and also display the grouped data in the common "stem." Boxplots display extreme values with dots, three quartiles with boxes, and may show a distance of 1.5 times the interquartile range with whiskers.

The statistical properties of commonly used distribution plots can be found in many references (Chambers, Cleveland, Kleiner, and Tukey 1983). Because each method has its own strengths and weaknesses, no single method works best for all data. In fact, due to the lack of flexibility in many packaged programs, even a collection of all standard methods has its limitations. Table 1 summarizes the features of four types of plots, namely, histogram/frequency polygon, 1-D scatterplot, stem-and-leaf plot, and boxplot, by indicating whether each plot provides a graphical presentation of selected summary statistics. Table 2 lists the strengths of these distribution plots, and when they can be used most appropriately.

In view of the limitations of standard distribution plots, our primary goal was to develop a versatile graphical tool with the following attributes: 1) suitable for displaying individual and grouped information, 2) applicable to continuous and discrete data, 3) providing innovative and effective ways for presenting data, and 4) producing standard and customized plots. In addition, we also tried to provide a tool with desirable features such as user friendliness, transportability between different machines and operating systems, and the ability to produce presentation-quality graphs.

This paper describes the features and interpretation of the BLiP plot, a function written in the S language (Becker, Chambers, and Wilks 1988; S-Plus Reference Manual 1991) that possesses these attributes. Section 2 delineates the BLiP plot with its arguments and capabilities. Examples are given to illustrate the uses of the variable-width plot, a unique feature of the BLiP plot. Section 3 gives two examples of the uses of the BLiP plot in facilitating data analysis. These examples illustrate the advantages of the BLiP plot over the standard plots. The interpretations of these graphs are also given. Section 4 discusses the trends in statistical graphics in presenting the 1-D distribution, and points out some future directions.

2. THE BLiP PLOT

The BLiP plot uses boxes, lines, and points as basic building blocks. Boxes and lines are used to display grouped information, while lines and points are used to present individual data values. The program is available as a function written in the S language. The source code of the program and the help file are available from the authors, and can also be retrieved through the S Archive in
Table 1. Capability of Commonly Used One-Dimensional Distribution Plots for Presentation of Selected Distribution Characteristics

<table>
<thead>
<tr>
<th>Distribution characteristics</th>
<th>Histogram/ freq. polyg.</th>
<th>1-D scatterplot</th>
<th>Stem-and-leaf plot</th>
<th>Boxplot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mean</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>S.D. or S.E.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Median, lower, and upper quartiles</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Modes</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Skewness</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Frequency</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

* The mean is not typically shown in the boxplot, and the median and the lower and upper quartiles are not shown in the stem-and-leaf plot, but these features can be added.
* The S.E. of the median can be shown in the notched boxplot.
* The mode and skewness are displayed explicitly in the stacking 1-D scatterplot and implicitly in the jittered 1-D scatterplot.

the StatLib (http://lib.stat.cmu.edu) or anonymous ftp by downloading the file /pub/S/blip.shar.Z from the Internet node odin.mdacc.tmc.edu (143.111.62.32). The BLIP function with its calling arguments is given below. Although 34 options are available for customizing the BLIP plot, a default value is provided for each one. The only required field is the name(s) of the variable(s) to be plotted (represented below by the function argument "..."). Detailed instructions for using these options can be found in the help file that comes with the program.

```
blip(..., std.plot, graph.region, graph.type, graph.width, graph.uniform, nclass, breaks, box.p, box.p.bar, box.label.on, box.label.cex, box.label.above, box.density, box.angle, box.col, line.p, line.sd, line.se, line.width, line.col, mean.pch, mean.cex, mean.col, point.pattern, point.type, point.pch, point.cex, point.col, xlim, axes, xaxt, xlab, ylab, seed)
```

One easy way for the first-time user to begin is calling blip() with the option std.plot = "histogram," "polygon," or "boxplot." The standard histogram, frequency polygon, or boxplot will be produced. The user can then start to experiment with the features of the BLIP plot by changing the default options.

A unique feature of the BLIP plot is its capability of drawing the distribution plot with variable graph width. When the graph.width = "v" the width at any particular

Table 2. Appropriate Applications of the Commonly Used One-Dimensional Distribution Plots

<table>
<thead>
<tr>
<th>Suitable for</th>
<th>Histogram/ freq. polyg.</th>
<th>1-D scatterplot</th>
<th>Stem-and-leaf plot</th>
<th>Boxplot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete data</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Continuous data</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Small sets (n ≤ 30)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mod. (30 &lt; n &lt; 100)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Large sets (n ≥ 100)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Outlier detection</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Gap detection</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Identifying each individual data value</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

* Stem-and-leaf plots can be used for continuous data after the data are discretized properly.
* Histograms and frequency polygons can detect outliers and gaps in the data with a suitable choice of interval width.
* Stem-and-leaf plots can only identify individual data values to a certain number of digits.

Figure 1. Variable-Width Jittered Plot.

\[
\hat{f}(x) = (\hat{F}(x + h/2) - \hat{F}(x - h/2))/h
\]

where \(\hat{F}(\cdot)\) is the empirical cumulative density estimation with linear interpolation and \(h = \text{range}/nclass\) where nclass is the smoothing parameter corresponding to the number of intervals used to divide the range of the variable. The central-difference estimator can be replaced by other estimators such as the kernel density estimator, if desired. The variable graph width is particularly useful in revealing the shape of the distribution when the point.pattern equals "jittered," "vertical-bar," or "max-range."

Figure 1a illustrates the variable-width jittered plot for the variable \(X_1\), which contains 500 random deviates generated from the standard normal distribution. Like the standard fixed-width jittered plot, the horizontal position of each point marks the actual value of each observation, and the vertical position is the result of random jittering for separating the ties. But unlike the standard jittered plot in which the vertical range of jittering is fixed, the variable-width plot varies the vertical range (i.e., width) to reflect the density estimation at each location. The variable-width jittered plot is effective, not only in revealing the shape of the distribution, but also in displaying the exact value of each single data point. The graph shows a nice, symmetrical, bell-shaped distribution with a center location around 0. It also indicates that there are 12 points above 1.96 and 15 points below \(-1.96\), with a range from \(-3\) to \(3.5\). We
expect that, on the average, there are $500 \times 2.5\% = 12.5$ observations less than $-1.96$ or greater than $1.96$. Therefore, Figure 1a illustrates that $X_1$ is generated from the standard normal distribution.

Figure 1b shows the distribution of $X_2$, which consists of $X_1$ plus 100 random numbers from the normal distribution with mean $= 4$ and variance $= 0.25$. The bimodal distribution as well as the value of each data point are clearly illustrated. The variable-width jittered plot combines the strength of the histogram and various forms of the 1-D scatterplot. The primary advantage of the variable-width jitted plot is that the shape of the distribution does not depend on arbitrary grouping of the data, as the shape of the histogram does. Moreover, users can vary the smoothing parameter to best display the distribution of the data and adjust the range of jittering. Note that on the vertical axes of Figure 1a, b, and similar variable-width plots to be shown later, no other values except the location of 0 are marked. This is because our main purpose is to display the shape of the distribution. Therefore, instead of computing the precise value of the density function, we only need to compute and plot the values proportional to the density estimation. With the variable-width options bplot() can also produce the "histplot" and "vaseplot" suggested by Benjamini (1988).

In addition to the 13 arguments for controlling the general graphical appearance, bplot() also supplies a rich set of eight arguments for boxes (box.pch, box.pch, box.pch, box.label.on, box.label.cex, box.label.label, box.angle, box.col), five arguments for lines (line.p, line.sd, line.se, line.width, line.col), and eight arguments for points (mean.pch, mean.cex, mean.col, point.pattern, point.type, point.pch, point.cex, point.col). One distinct advantage of using the BLIP plot is that bplot() provides users the freedom of manipulating the essential graphical elements. BLIP plot contains a set of tools for generating standard 1-D distribution plots as well as many innovative and customized plots. The following section gives two more examples to illustrate the main features of the BLIP plot.

3. TWO EXAMPLES OF DATA ANALYSIS

Selected key features of the BLIP plot are demonstrated in the following two examples. All plots are generated by bplot() except the stem-and-leaf plot.

**Example 1: Ozone Data.** The ozone data comprise 135 daily maximum ozone concentrations (in ppb or parts per billion) at ground level recorded between May 1 and September 30, 1974 in Stamford, CT for monitoring environmental pollution. Federal standards require that ozone concentration should not exceed 120 ppb more than one day per year, and an ozone concentration of 200 ppb is regarded as heavily polluted (Chambers et al. 1983). The data contain readings ranging from 14 to 240 with some ties. The 25th, 50th (median), and 75th percentiles are 48, 80, and 120.5 based on linear interpolation of order statistics. The mean is 89.8, and the standard deviation is 52.3. Various standard distribution plots of the data are shown in Figure 2.

Figure 2a shows the stem-and-leaf plot, with the stem in units of 10 and the leaf in units of 1. There are 34 days (25.2%) in which the ozone concentration was greater than 120 ppb. The plot is given in a horizontal format rather than the usual vertical format to facilitate comparison with

![Figure 2. Standard Distribution Plots for 135 Ozone Readings (in Parts per Billion). The data were recorded in Stamford, CT, from May 1 to September 30, 1974 (cf. Chambers et al. 1983).](image-url)
other figures. Figure 2b presents the histogram. The BLiP plot can display frequency counts of each interval at the top of the corresponding bars. The mean ± one standard deviation (89.8 ± 52.3) is shown under the histogram, with the mean marked by an “X.” The boxplot in Figure 2c summarizes the middle 50% of the data in boxes outlined by the 25th, 50th, and 75th percentiles. The whiskers may spread out from the 25th and 75th percentiles to the lower and upper ends of the data distribution, respectively, except when the distribution contains outliers. The length of the whisker can be as long as 1.5 times the interquartile range, but the whiskers do not extend beyond the data range. Two outliers (with values 230 and 240, which are larger than the upper quartile + 1.5× interquartile range = 229.25) can also be found on the right end of the whisker. A miniature version of the bar-code type 1-D scatterplot produced by specifying point.pattern = “vertical-bar” is given under the boxplot to indicate the position of each individual data point. Both Figure 2b and c are composite plots that can be easily produced in S by specifying par(new = T) and choosing the appropriate graph.region.

The shape of the distribution is characterized by the “density trace” (Chambers et al. 1983) in Figure 2d, shown with the default smoothing parameter. The density trace is formed by connecting the density estimation at each observed value with straight lines. Figure 2e shows the stacking 1-D scatterplot, which looks similar to the histogram, but without any grouping of the data. Tied observations are stacked up from the baseline with equal increments. For example, the mode of the raw data is at 38, where eight observations were recorded and can be easily identified by counting the number of points along the vertical axis. This information can also be found in Figure 2a, but is clearly displayed in Figure 2e. The fixed-width jittered 1-D scatter plot in Figure 2f breaks the ties by randomly jittering the points along the vertical axis. Although the exact value of each observation is precisely marked on the horizontal axis, the distribution of the data cannot be seen easily from the standard fixed-width jittered plot. In summary, Figure 2a–e are useful in identifying the distribution of the data, while Figure 2a, e, f, and the lower part of c are helpful in identifying each individual data value. All plots indicate that the data are skewed to the right, with the bulk of the data points lying between 50 and 120.

Customized BLiP plots using variable graph-width are shown in Figure 3. Figures in the upper row (a and b) use the baseline as the reference line (graph.type = “b”), while figures in the bottom row (c and d) are symmetrical about the center line (graph.type = “c”). Figure 3a is the variable-width version of the bar-code 1-D scatter plot. Each vertical bar indicates the exact position of each data point along the horizontal axis. The height of the bar corresponds to the density estimate at that value. When this plot is compared to the lower part of Figure 2c, the distribution of the data can be seen much more easily from the variable-width plot. Similarly, comparing Figure 3c to Figure 2f shows that the distribution information is embedded implicitly in the fixed-width plot, but is explicitly shown in the variable-width plot.

The choice of using dots or lines to present the data is the user’s. Dots are useful in separating the ties, while lines can offer better visual perception in marking the data values. Figure 3b shows the location of deciles instead of each individual data point. The height at each decile, again, is proportional to the density estimation at that value. Straight line segments are used to connect the vertical bars at the deciles. The figure shows that 80% of the data lies between 30 and 170, which correspond approximately to the 10th and the 90th percentiles. Figure 3d indicates the percentiles with 5% increments (0, 5, 10, 15%, . . . , etc.), but uses the center line as the reference line. The figure shows that an ozone concentration of 120 corresponds approximately to the 75th percentile. The percentile plot can be useful in comparing the distributions among different groups or in setting up standard thresholds for air pollution control, for example.

There is no definitive answer on whether the baseline or the center line is preferable as the reference line. Generally speaking, it is easier to identify the distribution produced by graph.type = “b” (baseline-referenced graphs) because the graph is similar to the usual density plot. On the other hand, graph.type = “c” (center-line-referenced graphs) may be better when comparing graphs because the symmetrical graphs are more suitable for side-by-side comparisons. Both types of graph are readily available in BLiP.

Example 2: Fisher’s Iris Data. Fisher’s iris data (Fisher 1936) contain four measurements on 50 flowers from each of three species of iris, namely, Setosa, Versicolor, and Virginica. We use only the sepal length (in centimeters) in our example to illustrate how the BLiP plot can be used for comparing the distributions among the species. Four plots of these data are shown in Figure 4. Figure 4a shows the fixed-width point plot with point.pattern = “stacking.” There is no grouping or jittering involved, nor is there a need to choose the smoothing parameter. All of the individual data points are plotted. In fact, each of the 50 data points
can be easily identified. Variable-width boxplots for the data are shown in Figure 4b. The boxes contain the middle 50% of the data as in the standard boxplot. However, instead of using 1.5 times the interquartile range, the whiskers in this plot extend from the 25th to the 2.5th percentile and from the 75th to the 97.5th percentile. Points outside the lines (2.5% at each end) are displayed as dots. The percentiles for drawing the boxes and lines can be changed according to the need. Extreme observations can be examined more carefully by noting the number and the location of the data points.

As can be seen, Virginica has the longest sepal length on average, followed by Versicolor and Setosa. Sepal lengths for Virginica are closer to those for Versicolor than to those for Setosa. Setosa has shorter sepal lengths and a tighter range than the other two. There are two observations from Virginica that fall below the lower 2.5th percentile. One observation is just below the lower 2.5th percentile, but the other one is farther away from other observations. The Virginica and Versicolor have the same minimum sepal length. However, the minimum sepal length of Virginica can be considered as an outlier. Figure 4c shows another version of the point plot and, again, the distribution of the data is better revealed in the variable-width plot. The mean ±2 S.D. is shown in Figure 4d with the means indicated by solid squares. Data outside the mean ±2 S.D. range are shown as dots. It is coincidental that several data points have values just outside the ±2 S.D. range, giving the appearance that the lines and dots are connected. Figure 4d is useful because it shows the locations of outliers on a plot commonly used in the sciences, the one that gives the location of the sample mean and the distance of 2 S.D. on each side of the mean. All plots show that Setosa has the shortest sepal length with the smallest S.D.

Figures 2–4 show that both the individual data value and summary statistics from the grouped information can be shown effectively by the BLiP plot. Instead of depending solely on the histogram or the boxplot, the BLiP plot provides applied statisticians a versatile tool for displaying data, and removes many limitations that exist in the available package programs.

4. DISCUSSION

Statistical graphics involve the effective use of both science and art. Tufte (1983) wrote: “Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency.” Standard statistical tools for displaying 1-D data offer precision. However, the “one size fits all” approach may not achieve the best clarity and efficiency for all varieties of data. This motivated us to develop a more flexible tool for displaying 1-D data. The BLiP plot allows users to choose any combination of boxes, lines, and points to show their data in the most effective way. In addition, commonly used distribution plots can also be easily generated. Standardization and versatility as well as ease of use and custom-made features are all combined under the same function call.

In the past two decades, especially after the microcomputer became popular in the 1980s, the trend in statistical graphics has been to bundle useful plot routines together in packaged programs. Hundreds of programs offering statistical graphs are available today. These programs typically can draw certain standard 1-D distribution plots easily, but allow very little room for users to customize their own plots. Because characteristics of data vary greatly from one study to another, and the purposes of data analysis may also vary, a more flexible graphical environment is needed. We took the first step to “unbundle” the graphical routines by developing the BLiP plot. The BLiP plot offers essential graphical ingredients, that is, boxes, lines, and points, for users to build their own plots. Advanced users can further modify the source code to suit their needs.

In the previous sections we introduced several new ways for presenting data more effectively by choosing different point patterns and using percentile plots and variable-width plots. Similar features can also be found in several commercial programs, such as the dotplot from SYSTAT (SYGRAPH 1993) and MINITAB (MINITAB 1996), the violin plot from NCSS (NCSS 1994), and the bar-code 1-D scatterplot from STATA (STATA 1992), but none are found together in one program. Only BLiP offers flexibility and versatility in handling all of these graphical elements. We hope that this is a new beginning of giving back to users the freedom to express their creativity and to enhance the clarity of presentation. There is always a danger of overuse and/or misuse of the power of statistical graphics, however. Users need to be aware of the strength and weakness of various approaches, and choose an unbiased way to present their data. Presentation bias can be reduced by showing data in several different formats, for example, by combining the standard 1-D distribution plots with various types of point plots or percentile plots.

Avenues to further improving graphical tools for presenting the data distribution are still wide open. For example, graphical user interface can be added to BLiP to facilitate its use. Interactive tools such as brushing can be developed.
A compiled version of BLiP in S (when the S compiler becomes available) or in other languages such as Visual Basic or Java will allow wider distribution and use.

[Received January 1995. Revised November 1996.]

REFERENCES


*STATA Reference Manual* (1992), College Station, TX: Stata Corporation.
