The Box-Percentile Plot

Warren W. Esty and Jeffrey D. Banfield

Abstract

A variant of the boxplot is proposed in which the sides contain the information of a percentile plot (which is equivalent to the empirical cumulative distribution function). Unlike boxplots, there is no question about how long to draw the whiskers, nor is there loss of information due to grouping. Side-by-side comparisons of distributions are especially effective. In spite of including more detail, the impact on statistically-untrained readers remains similar to that of traditional boxplots.

Keywords: boxplot, distribution, graphical data analysis, tree invasions.

1 Introduction

Consider the question of how to graphically communicate the information in several related sets of continuously-distributed data. This article introduces the box-percentile plot, a modified version of the boxplot that has two distinct advantages over previous versions. One is that there are no questions about how it should be drawn – the plot configuration is justified by the properties of empirical distributions and does not require arbitrary choices about box configuration or whisker length. The other is that it includes details which are useful to statistically-trained readers without weakening the impact on statistically-untrained readers – it is not necessary to choose between showing general behavior or detail (Tukey, 1977). The box-percentile plot can be used as both an exploratory tool and a space-saving method of publishing information in non-statistical scholarly articles which otherwise might take two or more graphs to convey.

Unlike the boxplot, which uses width only to emphasize the middle 50 per cent of the data, the box-percentile plot uses width to encode information about the distribution of the data over the entire range of data values. Figure 1 plots the highest points in 50 states and the heights of 219 volcanos (the same data used by Tukey (1977) for exhibit 5 of his Chapter 2) using box-percentile plots and standard boxplots. It is clear that the box-percentile plots convey the same graphical impression as the boxplots and contain additional information about the shape of the distributions. We will look at these plots in more detail in Section 4.2.

The idea behind constructing a box-percentile plot is simple. At any height the width of the irregular "box" is proportional to the percentile of that height, up to the 50th percentile, and above the 50th percentile the width is proportional to 100 minus the percentile. Thus, the width at any given height is proportional to the percent of observations that are more extreme in that direction. As in boxplots, the median, 25th, and 75th percentiles are marked with line segments across the box. Other percentiles may be emphasized if desired.

To illustrate the effectiveness of the additional information contained in the box-percentile plot (compared to the boxplot), consider the three artificial data sets shown as histograms in Figure 2. The distributions are clearly different. The first data set is from a standard normal distribution, the second data set is from a uniform distribution with an extreme outlier on each end, and the third data set comes from a tri-modal distribution. As shown in Figure 3, the boxplots of these data sets are indistinguishable while the box-percentile plots allow the observer to distinguish and identify the three distributions. We have obviously constructed this data to emphasize a strength of the box-percentile plot, however, in Section 4 we will use real data sets to further illustrate how the information provided by box-percentile plots can be used to solve problems that boxplots are unable to address.

2 Background

Tukey (1990) reminds us of the significance of impact, the distinction between prospecting and transfer of information, and the importance of recognizing the purpose to be served. Side-by-side boxplots indicate both the general magnitude of the observations in each set and permit rough comparisons between the sets. Even statistically-untrained readers can quickly grasp the basic distributional information displayed by the plot. The typical values and the general differences in the distribution of the values from set to set are easy to see. Then, only a small amount of statistical training is needed to be able to properly interpret the information that the quartiles and extreme values provide about the dispersion of the data. Thus, for communicating data, especially to scholarly but statistically-untrained readers, boxplots are often very effective and have become a standard graphical tool in many sciences.

Given the effectiveness of boxplots at communicating a few salient distributional details, the question naturally arises if boxplots can be modified to convey more information to the statistically-trained reader without confusing the picture for the uninitiated. Width has been used to encode information about sample size or confidence intervals (McGill, Tukey, and Larson, 1978). Shading has also been used to indicate confidence intervals (Benjamini, 1988). Another modification is to use width to indicate an estimated probability density. Unfortunately, the estimated probability density depends strongly on the method used to estimate it, so one set of data could yield widely varying plots (Benjamini, 1988). Variants of the boxplot without an actual box have also been proposed (Tufte, 1983; Tukey, 1977; Tukey, 1990). So far, none of these variants have found their way into common usage.

If the underlying principle of the basic, unmodified, boxplot method of visual presentation is accepted, there remains the question of precisely how to construct the plot (Frigge, Hoaglin, and Iglewicz, 1989). Boxplots emphasize a few key percentiles and mark them on the vertical scale. Horizontal marks are used primarily to emphasize the location of the middle 50 per cent of the data. They group data in the middle half into quartiles and group much of the rest into whiskers. Problems with boxplot construction are that the length of the whiskers and the treatment of extreme values are somewhat arbitrarily determined and no one definition has been accepted by all users. One way is to draw whiskers out to the 10th and 90th percentiles and mark each of the more extreme observations individually (Cleveland, 1985). In other versions the whiskers extend to the most extreme values (Moore and McCabe, 2003, p. 46; Tukey, 1977, p. 40; Tufte, 1983), or you may use discretion to chose

how many extreme values to indicate separately, in which case the whiskers extend out to them (Moore and McCabe, 2003, p. 47; Tukey, 1977, p. 41). Still another version (Tukey's "schematic plot", 1977, pp. 44, 47; Chambers, Cleveland, Kleiner, and Tukey, 1983) uses a far more complicated method to determine the lengths of the whiskers. The definitions of "steps" and "adjacent values" in this context are arbitrary. Lee and Tu (1997) developed a general graphical tool, the BLiP plot, which can produce many variations (34 options) on most standard distributional graphs, including boxplots. In the examples in their paper they use the .025 and .975 quantiles as the boundary for the whiskers in their boxplots. There is no general agreement as to what constitutes the basic boxplot. Thus when you see a boxplot you must scrutinize the accompanying text to determine how the plot has been constructed. The box-percentile plot eliminates the need to make these arbitrary choices.

Neither the boxplot nor the box-percentile plot is intended as a substitute for plots of estimated probability density functions. They do not plot densities and cannot be interpreted with the sophisticated (integral calculus) idea that area represents observations. The variable width BLiP plot (Lee and Tu, 1997) uses a simple central difference estimator of the density function to produce graphs that look similar to box-percentile plots but the encoded information and the interpretation is very different. As with all density estimates, the variable width BLiP plot requires a smoothing parameter and, depending on the choice of the parameter, can have several different representations of the same data. All distributional graphs that are based on density estimates entail arbitrary choices, and perhaps the most appealing feature of box-percentile plots is that they do not require such choices.

There are other types of diagrams which effectively display individual data sets without grouping and without questions about how they should be defined. In particular, percentile plots and empirical cumulative distribution functions display all the data and have straightforward definitions. For comparing several data sets, however, neither plot is suitable. If the graphs for several data sets are side-by-side the critical horizontal comparison is difficult to make, although summary lines can help (Cleveland, 1985, Fig. 3.19). Plotting several data sets on the same percentile or empirical cumulative distribution functions plot may lead to problems in detection and evaluation of differences between curves (Cleveland, 1985, Figs. 3.75 and 1.5).

3 The Box-Percentile Plot

The box-percentile plot combines the virtues of boxplots (ease of interpretation, ability to compare several data sets simultaneously) with those of percentile plots (display all the data, no arbitrary choices in construction). The idea is to use the width (as in the boxplot) to emphasize the middle of the data and to continue to use width (but not in an arbitrary fashion) to give less emphasis to the more extreme data (as whiskers and outliers are given less emphasis in boxplots). Thus the box-percentile plot "boxes" are wide in the middle (like boxplot boxes), narrow away from the middle (as are whiskers) and very narrow at the extremes. Unlike boxplots, the width contains precise information about the distribution of the data. They contain all the information of percentile plots and permit an easy and accurate assessment of symmetry. Also, since the data are not grouped, grouping can never conceal significant information as it would in some examples (Tukey's (1977) weight of nitrogen example, p. 50).

Let the number of observations be n and the observed values be ordered lowest-to-highest as $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$. Each y-value is plotted as a distinct point, so no information is lost. Let the desired maximum width of the box be w. If the data is a random sample, under rather general conditions it can be proved that the expected probability between the i^{th} and $(i+1)^{st}$ order statistic is $1/(n+1)$, so in percentile plots the data-point $y_{(k)}$ is marked at height $y_{(k)}$ above the horizontal coordinate $k/(n+1)$. In box-percentile plots the data-point $y_{(k)}$ is marked at height $y_{(k)}$ at distance $kw/(n+1)$ on either side of a vertical axis of symmetry, if $y_{(k)}$ is less than or equal to the median. If $y_{(k)}$ is greater than the median, it is plotted at height $y_{(k)}$ at distance $(n+1-k)w/(n+1)$ on either side of the vertical axis of symmetry. If the data is a population rather than a sample, then division by $n+1$ is inappropriate. In that case $y_{(k)}$ could be plotted at distance $(k-1)w/(n-1)$ from the axis for $y_{(k)}$ less than or equal to the median and at distance $(n-k)w/(n-1)$ if $y_{(k)}$ is above the median.

4 Examples

4.1 Simulated Data

Box-percentile plots not only facilitate comparison of a few key percentiles, as is the case with boxplots, they also permit comparison of complete distributions. However, if one wishes to ignore the additional information contained in box-percentile plots they can be used in the same manner as boxplots with no additional effort on the part of the reader. On the other hand, if one wishes to use the additional information provided by box-percentile plots then a small amount of training is needed. This subsection uses simulated data to help familiarize the reader with some of the common patterns that emerge in box-percentile plots. The following subsections will illustrate the use of box-percentile plots with real data sets.

Figure 2 shows the histograms of three simulated datasets and Figure 3 shows the corresponding boxplots and box-percentile plots. There are 300 observations in each sample so the plots have settled down to a fairly stable shape.

The first data set in Figure 2 is from a normal distribution. The corresponding box-percentile plot in Figure 3 shows a typical box-percentile plot for normal data. There is a single mode (no flat vertical lines) that occurs at the median. The plot is vertically symmetric about the median and the sides of the box are concave.

The second data set in Figure 2 is from a uniform distribution with two single outliers, one in each direction. The corresponding box-percentile plot in Figure 3 indicates there are outliers and the main body of the data is uniformly distributed because the sides are straight.

Outliers cause a long, thin line leading from the main body of the plot to the outliers. If there are several outliers in one direction, the "arm" of the box-percentile plot may have some width but they are still easily identifiable as unusual values that may not belong to the main body of data. Figure 4 shows a normal data set with several outliers in one direction. Notice how the main body of the box-percentile plot has the typical shape of a normal distribution while the narrow arm leading to the outliers easily identifies that set of points as outliers, without an arbitrary definition of "outlier." With regular boxplots, if some of

the observations are regarded as outliers, extending the whiskers out to the extreme values may give a misleading impression of where the data lie so outliers are generally marked individually. This is not a problem with box-percentile plots. However, if one wishes to define outliers and then emphasize them, the box-percentile plot can be modified to do this.

Regardless of the outliers, the box-percentile plot of the uniform data with outliers in Figure 3 still has the characteristic "diamond" shape of the uniform distribution (the percentile plot of a uniform distribution is linear). Compare the box-percentile plot of the uniform data with outliers from Figure 3 with the box-percentile plot of normal data with outliers in Figure 4. It is easy to distinguish between the distributions of the main body of the data, even in the presence of outliers.

The tri-modal box-percentile plot in Figure 3 illustrates the typical feature of a multi-modal distribution, vertical lines in the outline of the box. The "valleys" between modes have few observations relative to the "peaks", so there is little change in the percentiles in those regions which translates into flat, near vertical lines.

Figure 5 shows the box-percentile plot of a χ^2 data set. This illustrates the typical pattern of a box-percentile plot for skewed data. Compare Figure 5 to Figure 4 and it is easy to see the difference between a data set that is skewed and one that has outliers.

4.2 Volcanos

Figure 1 shows box-percentile plots and boxplots for two datasets, the highest point in each of the 50 states and the heights of 219 volcanos. In the boxplots it appears that the states data is skewed toward the higher values and there are a few outliers in the volcano data. The boxplots give no detailed information, other than location, about how these two datasets are related.

The box-percentile plots provide a far more informative view of the data. We can see from the box-percentile plot that the states data is bimodal rather than skewed. The shorter heights appear somewhat uniformly distributed between 0 and about 8000 feet. There is another group of states with heights between 13000 and 15000 feet and a single outlier (Alaska) at 21000 feet. Note that the boxplot for the states data gives the *impression* that there may be several states with maximum heights between 8000 and 11000 feet (the upper part of the box) while the box-percentile plot clearly shows that this is not the case.

4.3 Tree Invasions

Box-percentile plots provide a means to easily compare distributions and, as any good exploratory tool should, they can lead researchers to ask new questions about the data. A tree invasion is an encroachment of trees into a region where they have not traditionally grown. This is a serious problem in many parts of the world due to the loss of farming and grazing land. Figure 6 shows several data sets from a study of tree invasions in the upper Madison valley of Montana (Hansen, Wyckoff, and Banfield, 1995). A sample of trees were cored (a small part of the trunk is removed so the age of the tree can be determined by counting the tree rings) at several different invasion sites to determine when invasions occurred. Invasions show up as modes in the age distribution of the trees.

One of the more noticeable features of Figure 6 is the difference between the shapes of the plots which indicate different invasion histories. The trees at site 1 have a fairly uniform

age distribution indicating a fairly constant rate of germination. In contrast, site 2 had a small invasion (sudden increase in trees) in the early 1900's while site 3 had a strong invasion during the mid to late 1930's. Sites 5 and 6 have similarly shaped distributions that start at different times. Could there be something about the morphology of the land that could cause similar patterns? Sites 3 and 4 have similiar distributions until the strong invasion at site 3 in the late 1930's during which time site 4 had virtually no increase in trees germinating. Is there a difference between the fire or grazing histories of these two sites during that period? Also, the shape of the distributions for sites 3 and 4 after 1950 is different, the rate of germination at site 3 is smaller than that at site 4 (the width of the box is not changing as fast for site 3), could this indicate the sudden influx of trees at site 3 during the 1930's is suppressing the germination of new trees in the years that follow?

Box-percentile plots are, of course, not the only graphical tool that should be used in the exploration and analysis of this data. They are, however, a powerful tool for an initial view of what is occuring, in terms of tree invasions, over a large geographic area. In the hands of geographers, familiar with the potential causes of tree invasions, climate data, and grazing histories of the sites, the differences in the age distributions that can be seen and compared across sites using box-percentile plots provide insights and comparisons that are simply not available with other methods.

5 The Box-Percentile Plot Code

An R function has been written to implement box-percentile plots. The code, which is short and relatively easy to understand, is available as an ascii file at the Journal of Statistical Software (www.jstatsoft.org). A version of bpplot (based on an earlier version of our code and modified by Frank Harrell) is also available from the R archive (`lib.stat.cmu.edu/R/CRAN/`) in the package `Hmisc`. Box-percentile plots may also be created in Rweb (`www.math.montana.edu/Rweb`), a Web based interface to R (Banfield, 1999), just attach the box-percentile plot function with the command:

```
attach("/export/faculty/umsfjban/bpplot.rda")
```

The box-percentile R function, `bpplot`, has behavior similar to that of the boxplot function. It will accept an arbitrary number of variables (or a single list of variables) and it allows you to title the plot and label the axes. Besides the list of variables to be plotted, there are five named arguments that may be supplied:

- `names` ... a character vector to label the individual plots
- `main` ... a character string to title the plot (default is "Box-Percentile Plot")
- `xlab` ... a character string to label the x-axis (default is no label)
- `ylab` ... a character string to label the y-axis (default is "Percentiles")
- `population` ... a logical variable indicating whether or not the data represent a population (default is F). The last paragraph of Section 3 discusses how a box-percentile plot is calculated for a population.

The following code snippet is from the help page for the R function `boxplot` modified to compare boxplots with box-percentile plots. If you are using Rweb, uncomment the first line of code (remove the # symbol) to load the box-percentile plot function.

```
#attach("/export/faculty/umsfjban/bpplot.rda")

mat <- cbind(Uni05 = (1:100)/21, Norm = rnorm(100),
             T5 = rt(100, df = 5), Gam2 = rgamma(100, shape = 2))
boxplot(data.frame(mat), main = "Boxplots")
bpplot(data.frame(mat), main = "Box-Percentile Plots")
```

The following R code illustrates some of the options that may be used with `bpplot`.

```
x1 <- rnorm(100)
x2 <- runif(200, -1, 3)
x3 <- rchisq(150, 3)
alist <- list(x1, x2, x3)

bpplot(x1, x2, x3)
bpplot(alist, names=c("Normal", "Uniform", "Chi Squared"))
bpplot(Normal=x1, Uniform=x2, ChiSquared=x3)
bpplot(x1, x2, x3, xlab="Distributions",  ylab="Quantiles", population=T)
```

6 Conclusion

The box-percentile plot is not an all-purpose graph, but it does everything the boxplot does, and more, without being more difficult to interpret. And, fortunately, its construction is based on principles of mathematical statistics, not on arbitrary rules. There is no question about how long to draw the whiskers or how to plot the outliers. Thus this new type of graph not only has a visual impact which provides incisive comparisons, it presents the information in a statistically justified manner.

7 References

Banfield, J. D. (1999), "Rweb: Web-based Statistical Analysis" *Journal of Statistical Software,* 4.1, 1-15.

Benjamini, Y. (1988), "Opening the box of a boxplot." *The American Statistician,* 42.4, 257-262.

Chambers, J., W. Cleveland, B. Kleiner, and P. A. Tukey (1983) *Graphical Methods of Data Analysis,* Belmont, CA: Wadsworth International Group.

Cleveland, W. (1985) *The Elements of Graphing Data,* Monterey, CA: Wadsworth Advanced Books and Software.

Frigge, M., D. Hoaglin, and B. Iglewicz (1989) "Some implementations of the boxplot." *The American Statistician,* 43.1, 50-54.

Hansen, K., W. Wyckoff, and J. Banfield (1995) "Dynamics of Tree Invasions," *Forest and Conservation History,* 39.2, 66-76

Lee, J. Jack and Tu, Z. Nora (1997) "A Versatile One-Dimensional Distribution Plot: The BLiP Plot." *The American Statistician,* 51.4, 353-358.

McGill, R., J.W. Tukey, and W. Larsen (1978) "Variations of boxplots." *The American Statistician,* 32.1, 12-16.

Moore, D.S. and McCabe, G.P. (2003) *Introduction to the Practice of Statistics,* New York: W. H. Freeman and Company.

Tufte, E. (1983) *The Visual Display of Quantitative Information,* Cheshire, CT: Graphics Press.

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

Tukey, J.W. (1990) "Data-based graphics: Visual display in the decades to come." *Statistical Science,* 5.3. 327-339.

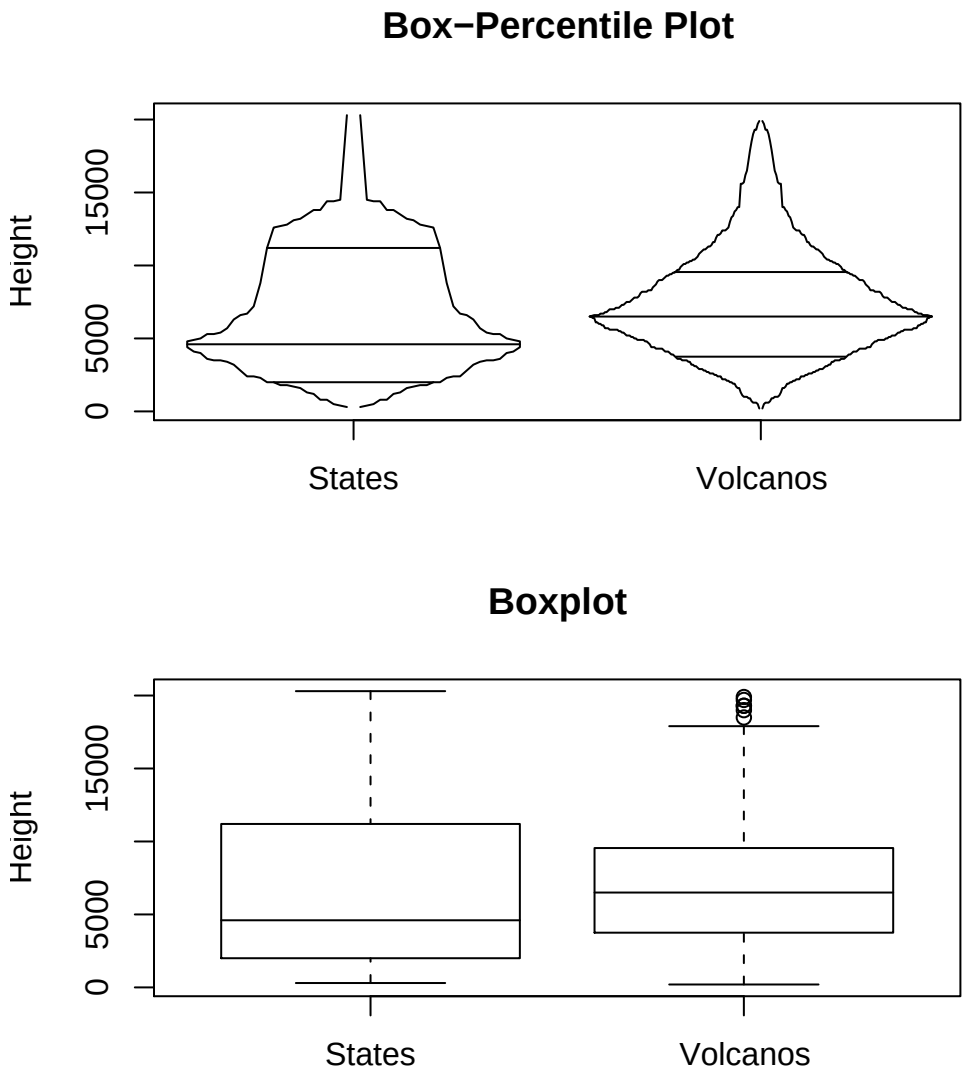

Figure 1: A comparison of Boxplots and Box-Percentile plots for the highest points in the fifty states and the heights of the 219 highest volcanos (Tukey (1977), Exhibit 2.5). The Box-Percentile plots provide more detailed information about the distribution of the data.

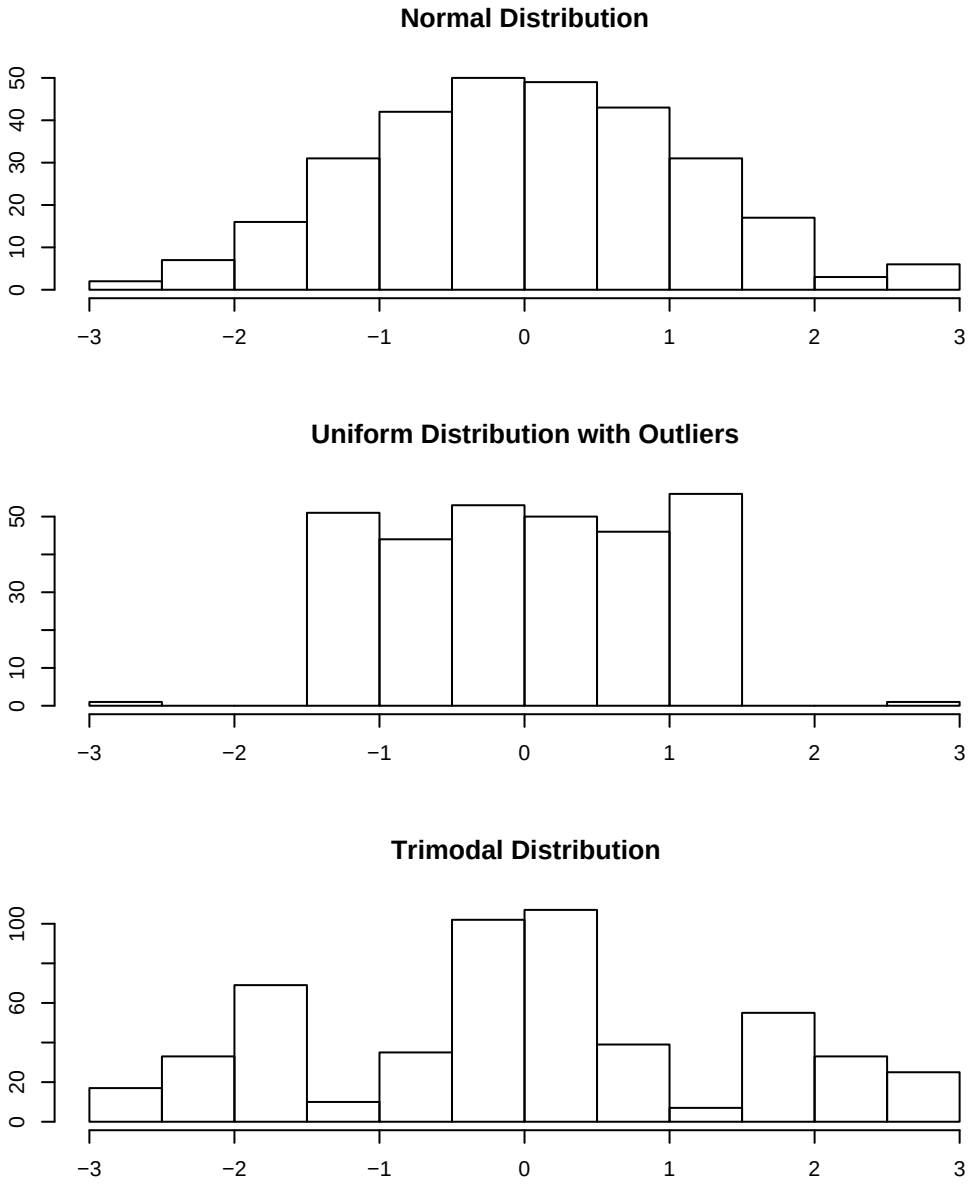

Figure 2: Histograms of three simulated data sets. Figure 3 uses these data sets to compare the effectiveness of boxplots and box-percentile plots in distinguishing between these three distinctly different distributions.

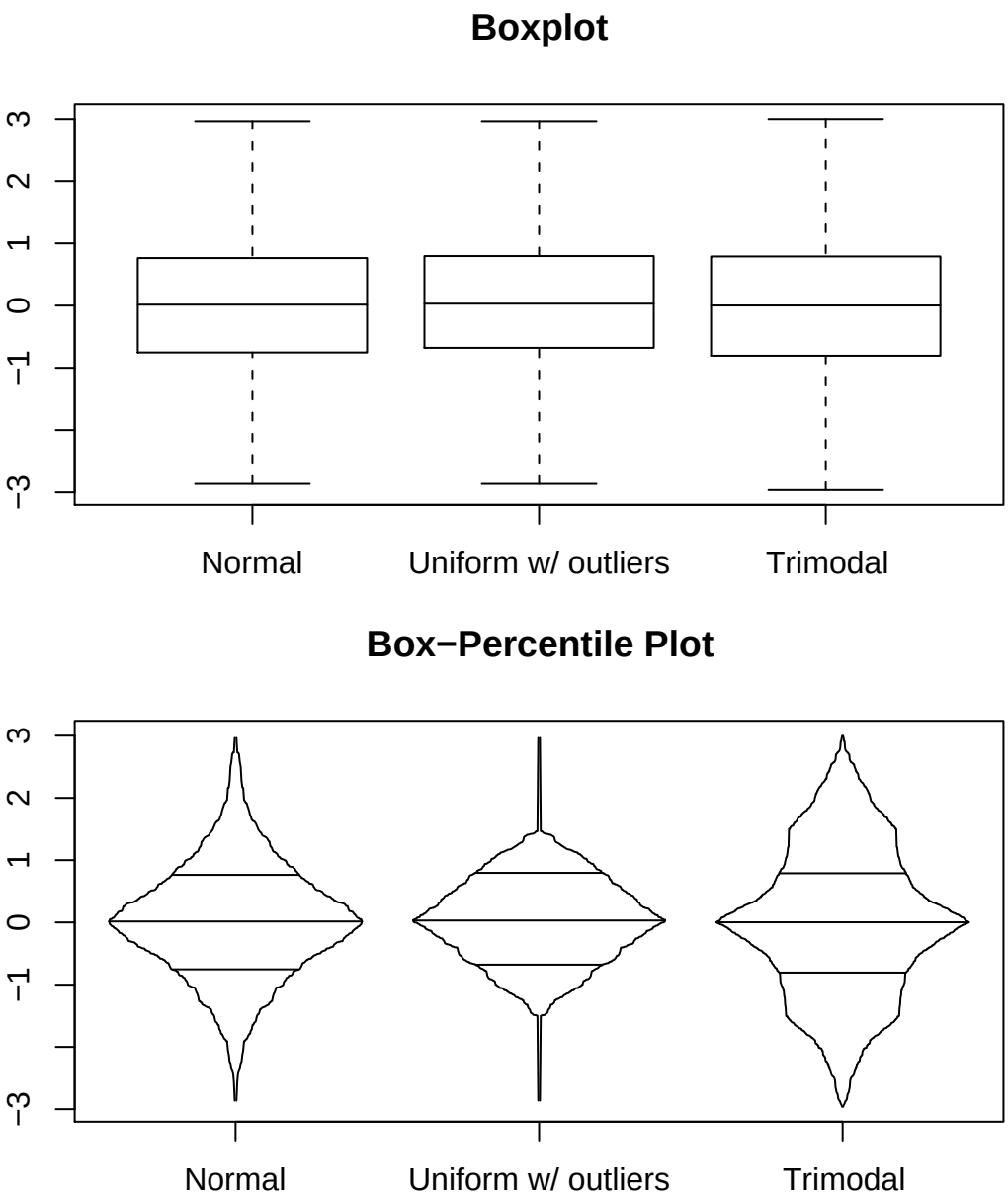

Figure 3: Boxplots and box-percentile plots for the data sets shown in Figure 2. Although all three boxplots are essentially identical, the box-percentile plots show the three data sets are very different.

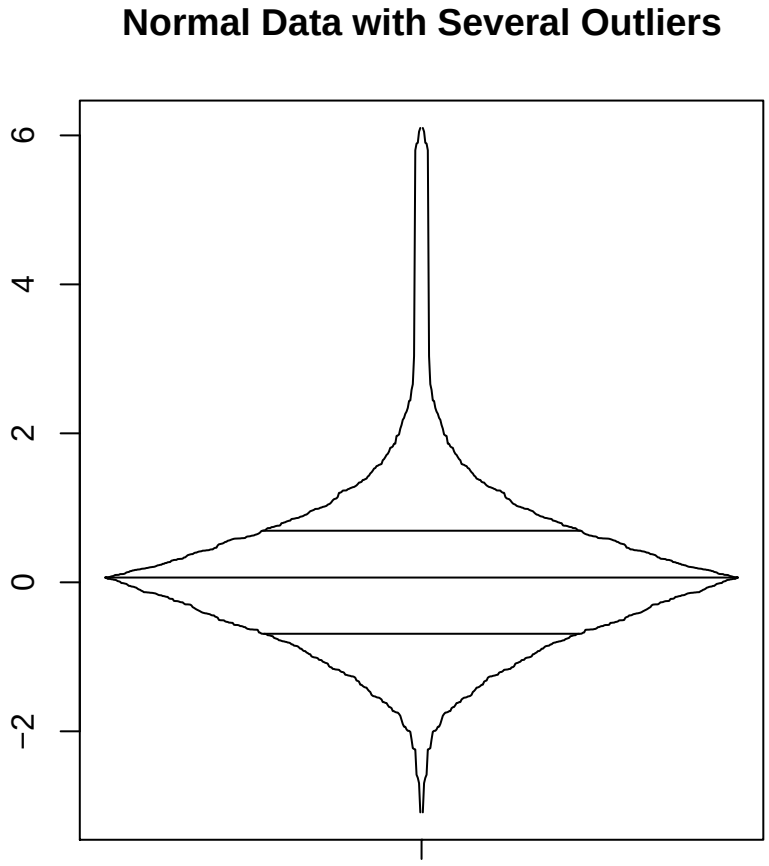

Figure 4: This plot shows the effect of a few outliers on the box-percentile plot for normal data. Compare this plot to the box-percentile plot for normal data shown in Figure 3.

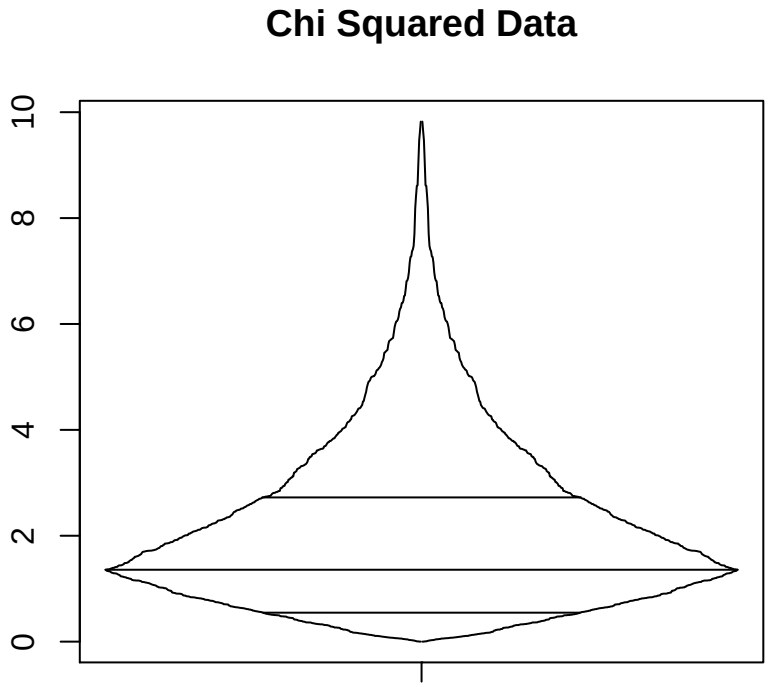

Figure 5: A box-percentile plot showing the typical pattern for skewed distributions. Compare this plot to the box-percentile plot for normal data with outliers shown in Figure 4.

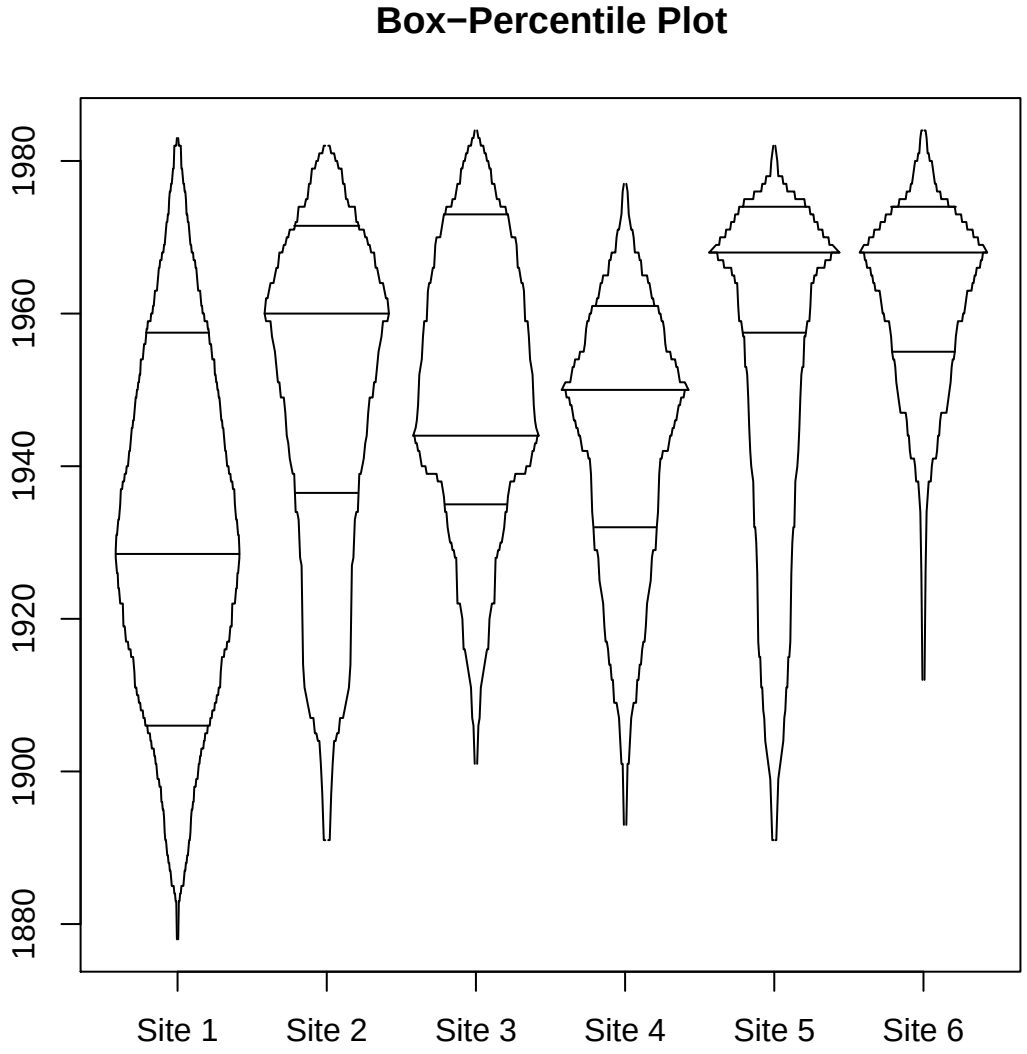

Figure 6: Box-percentile plots for tree invasions at six different sites in the Madison Valley of Montana. A tree invasion is encroachment of trees into a region where they have not traditionally grown. One of the more noticeable features of these box-percentile plots is the difference between the shapes of the plots which indicate different invasion histories.