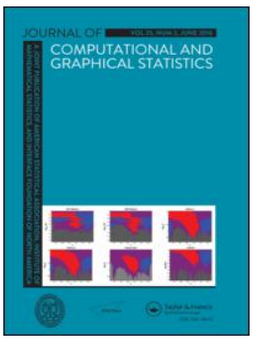


Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

# Letter-value plots: Boxplots for large data

Heike Hofmann, Hadley Wickham & Karen Kafadar

**To cite this article:** Heike Hofmann, Hadley Wickham & Karen Kafadar (2017): Letter-value plots: Boxplots for large data, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2017.1305277

**To link to this article:** http://dx.doi.org/10.1080/10618600.2017.1305277

# Letter-value plots: Boxplots for large data

Heike Hofmann
Department of Statistics, Iowa State University
and
Hadley Wickham
RStudio
and
Karen Kafadar
Department of Statistics, University of Virginia

March 6, 2017

**Abstract**

Boxplots (Tukey, 1977) are useful displays that convey rough information about the distribution of a variable. Boxplots were designed to be drawn by hand and work best for small data sets, where detailed estimates of tail behavior beyond the quartiles may not be trustworthy. Larger data sets afford more precise estimates of tail behavior, but boxplots do not take advantage of this precision, instead presenting large numbers of extreme, though not unexpected, observations.

Letter-value plots address this problem by including more detailed information about the tails using "letter values", an order statistic defined by Tukey. Boxplots display the first two letter values (the median and quartiles); letter-value plots display further letter values so far as they are reliable estimates of their corresponding quantiles. We illustrate letter-value plots with real data that demonstrate their usefulness for large data sets.

All graphics are created using the R package `lvplot`, and code and data are available in the supplementary materials.

*Keywords:* order statistics, quantiles, fourths, tail area, location depth.

## 1 Introduction

Boxplots (Tukey, 1970, 1972) are one of the few statistical graphics invented in the 20th century that have gained widespread adoption. Despite their widespread use, they are not altogether satisfactory, particularly for large data sets. Specifically, two problems arise with boxplots when applied to large data sets: (1) the number of outliers (observations beyond the whiskers) grows linearly with the sample size and (2) estimates of tail behavior are not displayed, despite the fact that larger sample sizes allow more reliable estimates further out into the tails. Boxplots also have problems for very small number of values — while it is possible to draw a boxplot based on fewer than five points, the information gained from the quartiles is highly suspect.

Figure 1 illustrates both problems of conventional boxplots for the square roots of taxi times of 469,968 U.S. domestic flights in January 2015 stratified by 15 U.S. airline carriers. Carriers are ordered by the total number of flights from left (largest) to right (smallest). The largest three carriers, WN (Southwest), DL (Delta), and EV (ExpressJet) show striking differences in the median taxi times: the median taxi time for Southwest of $4^2 = 16$ minutes is below the overall median of 21 minutes, while Delta has a median taxi time of 25 minutes.

The sample sizes in the fifteen carriers vary between 4,731 flights by VX (Virgin America) and over one million flights by WN (Southwest). Recall that roughly 0.7% of the observations for a sample from a Gaussian distribution can be expected to be labeled as "out" or "far out" in a boxplot (Hoaglin, Mosteller, Tukey 1983, p.34). Thus, with so many flights for each airline, the number of labeled outliers in Figure 1 is huge, far too many to investigate individually and to distinguish between expected extreme values and 'true outliers.

The construction of a boxplot is based on a set of order statistics called letter values, specifically the sample median M and the lower and upper fourths $L_F$ and $U_F$, called "hinges" in Tukey (1977). In *Exploratory Data Analysis*, Tukey (1977) recommended the use of order statistics as estimates for the fourths, specifically, the $(\lfloor n/2 \rfloor + 1)/2$-th and $(n + 1 - (\lfloor n/2 \rfloor + 1)/2)$-th order statistics for lower and upper fourths for a sample size of $n$.

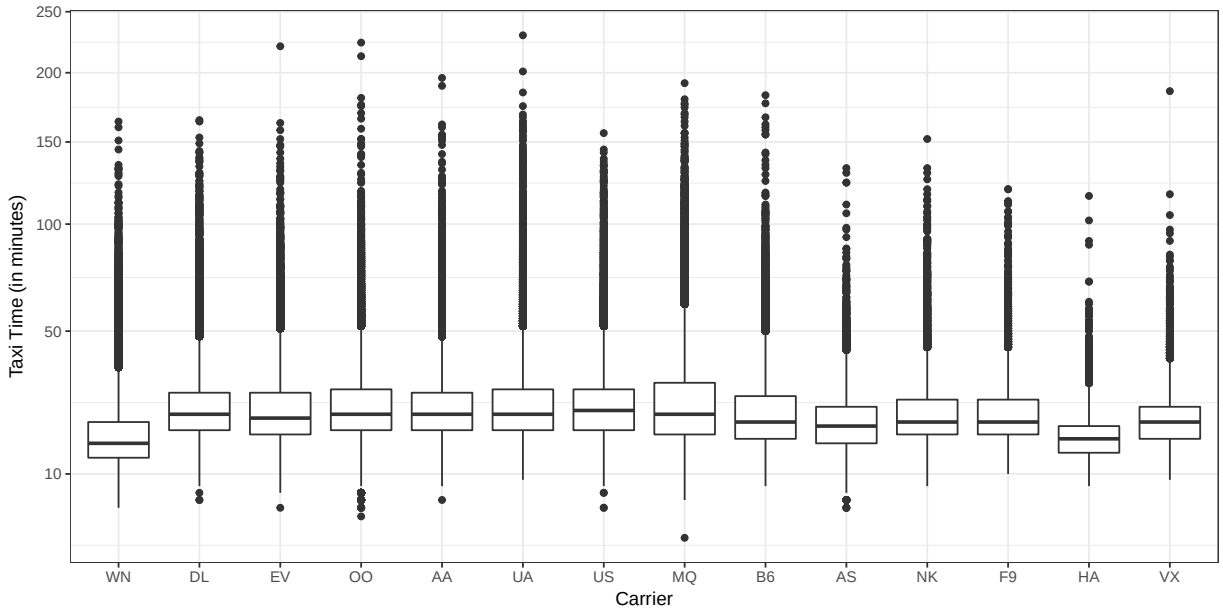



Figure 1: Boxplots of square root-transformed taxi in and out times grouped by airline carrier for flights in January 2015. Airline carriers are ordered left to right in decreasing number of flights served.

Similarly, the definition of "outside value", which we will denote here simply as "outlier", depends on the range between the fourths: an observation is labelled as an "outside value" and displayed individually if it lies at or beyond the inner fences (Tukey, 1977; Emerson and Strenio, 1983), defined as $[L_F - k(U_F - L_F), U_F + k(U_F - L_F)]$ where $L_F$ and $U_F$ denote the lower and upper fourths and typically $k = 1.5$.

Despite the name, these outliers may be either (a) genuine, but extreme, observations from the same distribution as the bulk of the data; or (b) true outliers, i.e. observations from a different distribution. The boxplot tends to display too many outliers, as judged by looking at boxplots of Gaussian data. There the expected number of outlier grows approximately linearly with the number of observations $n$, with about 0.7% of the points labeled as outliers (Hoaglin, 1983, cf.). The probability of getting at least one outlier for Gaussian data exceeds 30% for samples of size 50, and rises to 97% for samples of size 500 (Hoaglin et al., 1986, pg. 1148). The approach of Hoaglin and Iglewicz (1987) ("fixed outside rate") labels a fixed number of outliers using a rule based on the fourths. Although it avoids the linear dependence of the number of outliers on $n$, this approach also fails to display any interesting features in the tails. Large data sets permit many

more letter values that can be reliably estimated to provide more information about the tails.

Alternative displays have been proposed to illustrate further details of the distribution, such as multi-modality. The overview paper by Wickham and Stryjewski (2012) discusses these alternatives, in particular vase (Benjamini, 1988), violin (Hintze and Nelson, 1998), bean Kampstra (2008), box-percentile (Esty and Banfield, 2003), and high-density region (Hyndman, 1996) plots. These displays provide more detailed information about the distributions, through the use of non-parametric density estimates, which are especially useful for larger sample sizes. However, as Benjamini (1988) acknowledged, these displays depend on the specific estimation procedure (e.g., kernel density estimate) as well as on additional smoothing ("tuning") parameters. Thus, these displays can be different for the same data set, depending on the density estimate or smoothing parameters. As an initial exploratory visualization tool, this dependence on multiple tuning parameters is less than desirable.

Letter-value plots are a variation of boxplots that replace the whiskers—the values with the largest data depths within the fences—with a variable number of letter values, selected based on the uncertainty associated with each estimate, and hence the number of observations. Any values outside the most extreme letter value are displayed individually. These modifications reduce the number of outliers displayed for large data sets, and make letter-value plots useful over a much wider range of data sizes. Letter-value plots remain true to the spirit of boxplots by displaying only actual observations from the sample, and remaining free of tuning parameters.

Figure 2 shows a letter-value plot of the same taxi-ing times as Figure 1. Letter-value plots are better suited to show the skewed tails (even with the square-root transformations) and show far fewer labeled outliers. We will describe these plots in more detail in Section 3.

One consideration for letter-value plots involves the number of letter values to display, i.e. when to stop displaying letter values and start showing individual observations. Figure 2 shows only those letter values whose approximate "95% confidence intervals" do not overlap with the adjacent letter values. This varies based on the size of the group,

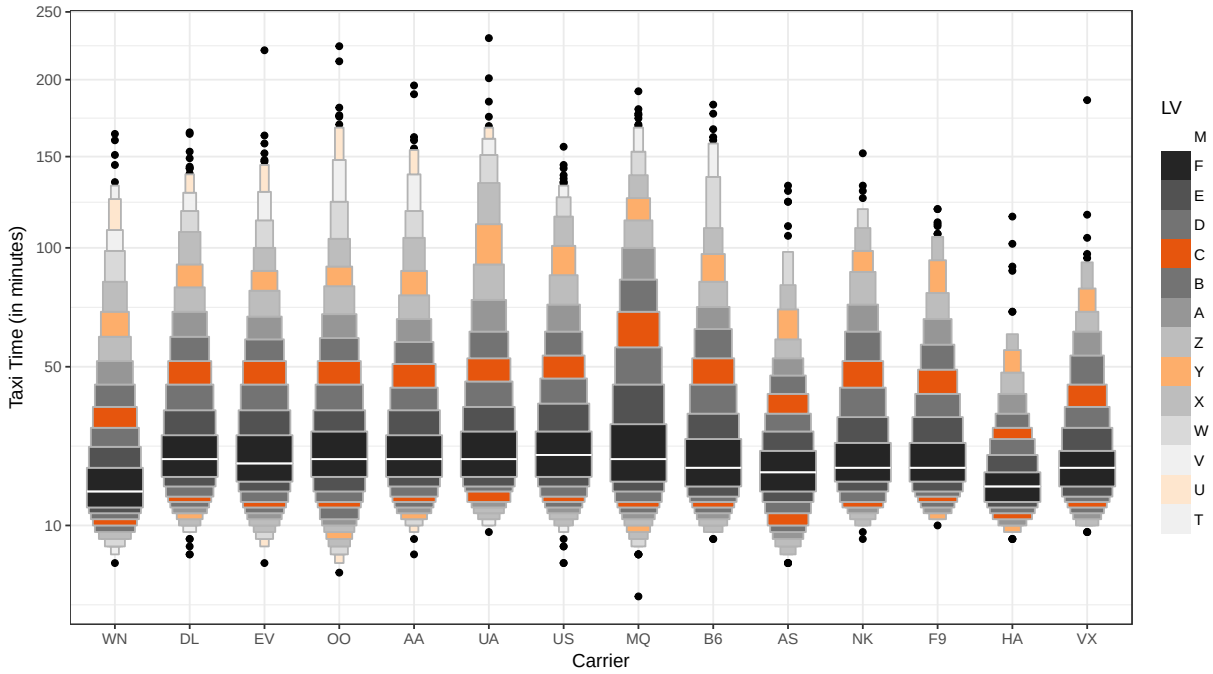



Figure 2: Letter-value plots of (sqrt-transformed) taxi time for 478,755 national flights in January 2015 by airline carriers. Colored bands among the grey-shades help to keep track of the same letter value across groups.

so different carriers show different letter values. Section 4 discusses three other rules to select the letter values based on the sample size. Some proposals for multivariate data and final discussion appears in Section 6.

Our implementation of letter-value plots is available in the R package, `lvplot`. The package also contains the data sets used here. The online supplementary material contains all code to reproduce the plots in this paper.

## 2 Letter values

Hoaglin and Tukey

Let $X_{(1)}$, ... $X_{(n)}$ denote the order statistics from a sample of size $n$. Per conventional notation, let $\lfloor y \rfloor$ and $\lceil y \rceil$ denote the largest integer not greater than $y$ and the smallest integer not less than $y$, respectively. The letter values are order statistics with specific depths, defined recursively starting with the median: the depth of the median, $d_1$ or $d_M$, of a sample of size $n$ is defined as $d_1 = (1+n)/2$; the depths of successive letter values (F = fourths, E = eighths, D = sixteenths, C = thirty-seconds, ...) are defined recursively as $d_i = (1 + \lfloor d_{i-1} \rfloor)/2$. We also will use the letter value itself as the subscript to the notation

for depth; e.g., both $d_2$ and $d_F$ denote the depth of the fourths.

If the depth is an integer plus $\frac{1}{2}$, the corresponding letter value is defined as the average of the two adjacent order statistics, $X_{(\lfloor d_i \rfloor)}$ and $X_{(\lceil d_i \rceil)}$ for the lower, and similarly for the upper letter value.

The $i^{th}$ lower and upper letter values ($LV_i$) are thus defined as $L_i = X_{(d_i)}$ and $U_i = X_{(n-d_i+1)}$. Because each depth is roughly half the previous depth, the letter values approximate the quantiles corresponding to tail areas of $2^{-i}$.

The "labeled outlier rule" for conventional boxplots relies on the fourths because the rule is then "unlikely to be adversely affected by extreme observations" and "to minimize the difficulties of masking" (Hoaglin et al., 1986, pg. 992). The breakdown point of these boxplots is 25%; i.e., only if 25% or more of the data values, all located in one of the tails, are contaminated, will the summary statistics and outlier identification change. This high breakdown is one of the valuable features of boxplots. In Section 5 we see that even though letter-value plots perform differently with highly contaminated data, they still show a set of valuable features.

The relatively low uncertainty in the fourths as estimates of the quartiles argues for using the fourths in the rule for labeling outliers: the standard deviation of the fourths in a Gaussian population equals roughly $[(0.25 \times 0.75)/(n\phi(\Phi^{-1}(0.25)))]^{1/2}\sigma = 1.362\sigma/\sqrt{n}$ or a 2-SD uncertainty of roughly $0.25\sigma$ for Gaussian samples of size 120 (David and Nagaraja, 2003). Estimates of the population quantiles beyond the quartiles, when based on order statistics, are increasingly variable; e.g., for the same $n = 120$ sample, the 2-SD uncertainty in the eighths (depth = 13) and sixteenths (depth = 7) is approximately $2 \times 1.607\sigma/\sqrt{n} = 0.29\sigma$ and $2 \times 1.968\sigma/\sqrt{n} = 0.36\sigma$, respectively. Table 1 shows these factors for the standard error formula, `SEfactor`, for the first eight letter values, as well as the relative increase in sample size needed for successive letter values to have the same uncertainty as the fourth. For example, the fourths in a sample of size 120 have a 2-SD uncertainty of $0.25\sigma$; we would need a sample of size $1.4\times120 = 168$ for the eighth to have this same level of uncertainty. For small samples, then, restricting attention to estimates of only the population median and quartiles, with some general indication of the tail length beyond the quartiles, is likely to be all the information that the data can

reliably support.

| LV | ideal tail area | rough % | odds ($2^i$) | SEfactor | n-equiv* |
|---|---|---|---|---|---|
| M | .50 | 50.0% | 2 | 1.25 | |
| F | .25 | 25.0% | 4 | 1.362633 | 1.0 |
| E | .125 | 12.5% | 8 | 1.60 | 1.4 |
| D | .0625 | 6.25% | 16 | 1.96 | 2.1 |
| C | .03125 | 3.13% | 32 | 2.47 | 3.3 |
| B | .015625 | 1.56% | 64 | 3.16 | 5.4 |
| A | .0078125 | 0.8% | 128 | 4.10 | 9.1 |
| Z | .00390625 | 0.4% | 256 | 5.37 | 15.6 |

Table 1: First 8 letter values. `Ideal tail area` is $2^{-i}$, $i = 1, ..., 8$. `rough%` rounds $2^{-i} \times 100\%$ to the first 1 or 2 nonzero digits. `odds` expresses tail area as `1 in` $2^i$. `SEfactor` gives the factor for the asymptotic standard error of the order statistic (from a Gaussian population, variance $\sigma^2$) corresponding to `tail area`. `n-equiv` = (`SEfactor`/1.362633)$^2$ which gives the factor of increase in sample size for the uncertainty in that letter value to be the same as that for the fourth.

Letter values are particularly useful for large data sets, because (a) much of the most valuable information, especially for inference purposes, is contained in the tails (cf. Winsor's principle, "All distributions are normal in the middle" (Tukey, 1960, pg. 457)); and (b) adjacent letter values have asymptotic correlation $\sqrt{1/2} = 0.707$ (Mosteller (1946) cited by Hoaglin (1983, pg. 51–52)). Thus, rather little information concerning tail behavior is lost by considering only the letter values.

## 3 Letter-value plots

Letter-value plots are based on the letter values, with a line displaying the median, and one box for each pair of adjacent letter values. Assuming we draw a vertical letter-value plot, the height of a box is fixed by the letter values, but we have three options for the width:

**1.** `linear`**:** make the width of each box inversely proportional to the letter value it represents, i.e. starting with the fourths, each successive box is one step narrower than the previous box, where the step size depends on the overall width of the plot and the deepest letter value shown.

**2.** `area`**:** make the *area* of each box proportion to the number of points in it. Because number of observations between letter values $LV_i$ and $LV_{i+1}$ is roughly $1/2^{i+1}$ of the total, the width of the $i$th box (counting from the median out) is inversely proportional to $2^{i+1}\,|LV_i - LV_{i+1}|$.

**3.** `height`**:** make the *width* of each box proportional to the number of points in it, i.e. the box of the $i$th letter value has a (relative) width of $2^{-i}$.

Area-adjusted widths ensure that the overall area is close to one, or more precisely, if $k$ letter values are shown, the overall area will be proportional to $1 - 2^{-(k+1)}$. This makes an area-adjusted letter-value plot a representation of the variable's density. Side-by-side versions of these plots show conditional densities, i.e while they do not explicitly contain the number of values (sample size is shown indirectly by the number of letter values chosen), the boxes of corresponding letter values have the same size. For options (1) and (3) boxes with matching heights (for horizontal plots) correspond to the same depths.

The box colors are chosen to aid comparison across groups. The innermost boxes are shaded heavily more heavily to represent greater data density, and every fifth box is given a contrasting hue. In Figure 2, observe the very compact distribution of taxi times for Hawaiian airlines (HA): the top of the second red band above the median indicates the seventh letter value located at about 50 minutes, so only $2^{-7} = 1/128$th of the values are greater than 50. This is much lower than the other carriers.

Beyond the most extreme box, all observations are identified individually. With this definition, the expected proportion of the outliers (roughly $1/2^i$) equals the expected proportion between this end and the end of the next bigger box (i.e., roughly $1/2^{i-1} - 1/2^i = (1/2^{i-1})(1 - 2^{-1}) = 1/2^i$). When the depth, $d_i$, reaches 1, the letter values are the extremes (minimum and maximum).

Letter-value plots for three different distributions are shown in Figure 3. Each panel displays a sample of 10,000 data points (left = standard Gaussian, middle = exponential with mean 1, right = standard uniform), first using the conventional boxplot (top row) and then with the proposed letter-value plot up to letter Y, corresponding roughly to tail area $2^{-9}$ = 1/512 (bottom row). Comparing the left (Gaussian) and right (Uniform) letter-value plot, overall *more heavily shaded* displays correspond to distributions with

*lighter* tails.

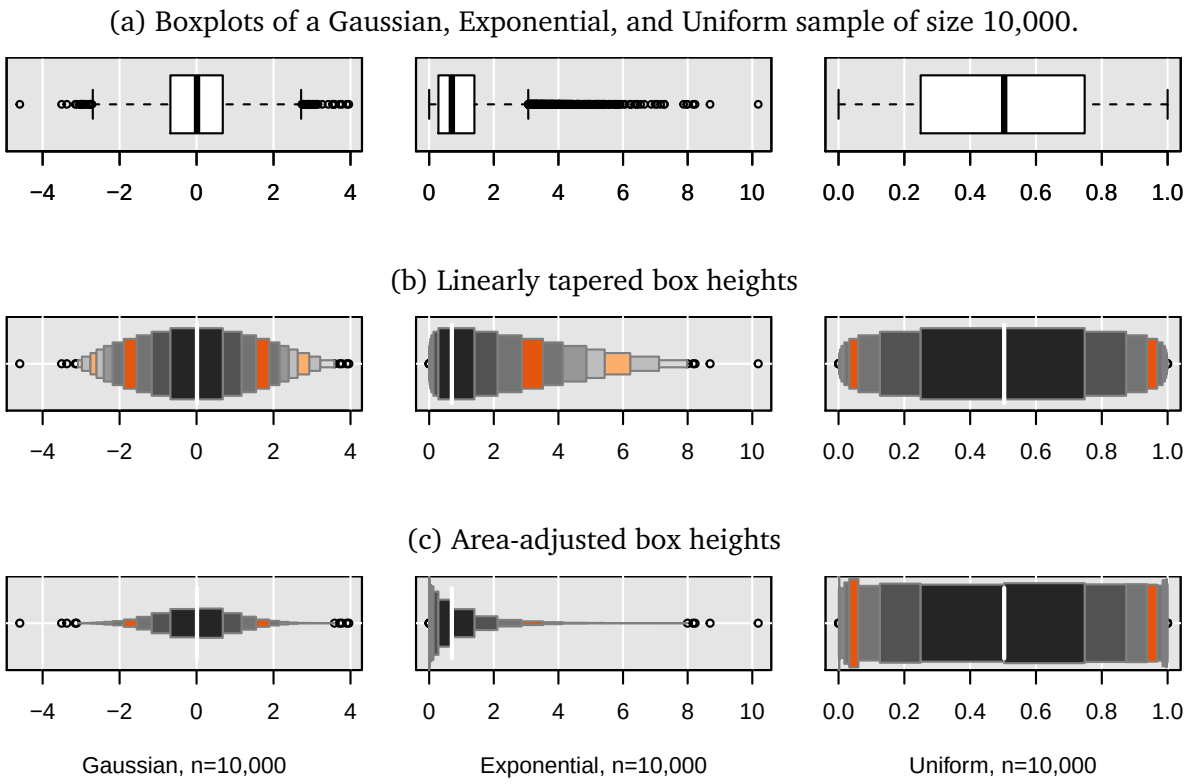



Figure 3: Standard boxplots (top row) and letter-value plots (middle and bottom row) for data from three different distributions. Each plot shows 10,000 data points. From left to right, samples come from $\mathrm{N}(0, 1)$, $\mathrm{Exp}(1)$, and $\mathrm{U}[0, 1]$.

## 4 How far out? — Stopping rules

We need a rule to determine the number of boxes to show in a letter-value plot, which will in turn determine the number of labeled outliers. In this section, we consider four proposals for such a rule. The first two rules are based on using a constant—either absolute or relative to sample size—the second two rules are based on 'trustworthiness' and make use of the standard error of the letter values involved:

(i) **Rule 5–8:** we can choose the extent of the letter-value plot display so that the last set of boxes encompasses all but the 5–8 most extreme observations, if we stop the letter-value plot display at $LV_k$ where

$$k = \lfloor \log_2 n \rfloor - 2. \tag{1}$$

Note that for small sample sizes the letter-value plot simplifies under this rule to a dotplot ($n < 8$) or a dotplot with a median ($8 \leq n < 16$).

Tukey identified 5–8 extreme points in many of the displays in *Exploratory Data Analysis*.

(ii) **Constant proportion:** an alternative criterion fixes the number of labeled outliers as a proportion of the overall sample size. Let $p$ denote this proportion, then we stop the letter-value plot display at $LV_{k_p}$ where

$$k_p = \lfloor \log_2 n \rfloor - \lfloor \log_2(np) \rfloor. \tag{2}$$

In effect, conventional boxplots use this rule for outlier identification, with $p = 0.007$. This criterion results in the same rule as the previous rule for the samples in Figure 3 ($n = 10,000$) when $p$ lies between 0.0004 and 0.0007 (0.04–0.07%), and in Figure 6 ($n = 3068$) when $p$ lies between 0.0014 and 0.0026 (0.14–0.26%).

(iii) **'Trustworthiness':** to assess the "trustworthiness" of the $k^{th}$ letter value we interpret the values as an estimate of the corresponding population quantile. We can "trust" a given letter value, if its approximate $100(1-\alpha)$% confidence does not overlap the subsequent letter value. This leads to a straightforward rule for $k$ given as (see below for a proof):

$$k_{1-\alpha} = \lfloor \log_2(n) \rfloor - \lfloor \log_2 \left(2z^2_{1-\alpha/2}\right) \rfloor. \tag{3}$$

Interestingly, using a 95% confidence interval, this rules leads to labeling 5–8 of the most extreme observations on each side, surprisingly consistent with many of the displays in Tukey (1977).

(iv) **Maximal standard error**: this fourth rule gives us the maximum letter value for which the standard error is fixed to be below a relative upper limit.

For a sample from a Normal distribution, the asymptotic standard error of the $i$th

letter value is given as

$$SE(LV_i) \approx \sigma\sqrt{p_i \cdot (1-p_i)/n}/\phi(\Phi^{-1}(p_i)) = (\text{SEfactor})\sigma/\sqrt{n}, \tag{4}$$

where $p_i = 2^{-i}$, $i = 1, ..., \log_2 n$, $\texttt{SEfactor} = \sqrt{p_i(1-p_i)}/\phi(\Phi^{-1}(p_i))$ and $\phi, \Phi$ are density and distribution of the standard Normal distribution. As sample size $n$ increases, we can increase the number of letter values shown while keeping the same standard error. Figure 4 gives an overview of the situation. We see that the logarithm of the sample size is approximately linear in the letter value. This rule is very similar to rule (1) when the desired uncertainty in the letter value does not exceed $0.25\sigma$.

*Proof.* for (iii): Consider the upper $k^{th}$ letter value, $LV_k$. We can view this value as the median between the extreme and the previous letter value, $LV_{k-1}$. An approximate $(1-\alpha)100\%$ confidence interval for the median of $m$ values (with $m > 10$) contains approximately $0.5\sqrt{m}z_{1-\alpha/2}$ observations on each side of the sample median (rounding to the nearest integer), where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard Gaussian distribution (David and Nagaraja, 2003, p.161).

The $k$th letter value is approximately the median of $2d_k$ values, its $(1-\alpha)100\%$ confidence interval therefore encompasses approximately $0.5\sqrt{2d_k}z_{1-\alpha/2}$ values on each side.

Between letter value $\text{LV}_k$ and $\text{LV}_{k+1}$ are approximately $d_{k+1}$ values. We therefore stop at letter value $k$, if

$$0.5\sqrt{2d_k}z_{1-\alpha/2} > d_{k+1}.$$

Because $2d_{k+1} \approx d_k$ this leads us to stop at $k$ when there are fewer than $2z^2_{1-\alpha/2}$ observations in the tails, i.e.

$$k_{1-\alpha} = \lfloor \log_2 n \rfloor - \lfloor \log_2(2z^2_{1-\alpha/2}) \rfloor.$$

□

The third stopping rule has the obvious advantage that it provides a simple, distribution-free solution. The sample size determines the number of letter values to be used, and therefore has only a small effect on the number of outliers. When $\alpha = 0.05$ (95% point-

wise confidence), the rule leads to showing only those letter values whose depths are at least 10 (i.e., labeling 5–8 observations on each side). No other distribution-related characteristic such as skewness or kurtosis, affects the rule.

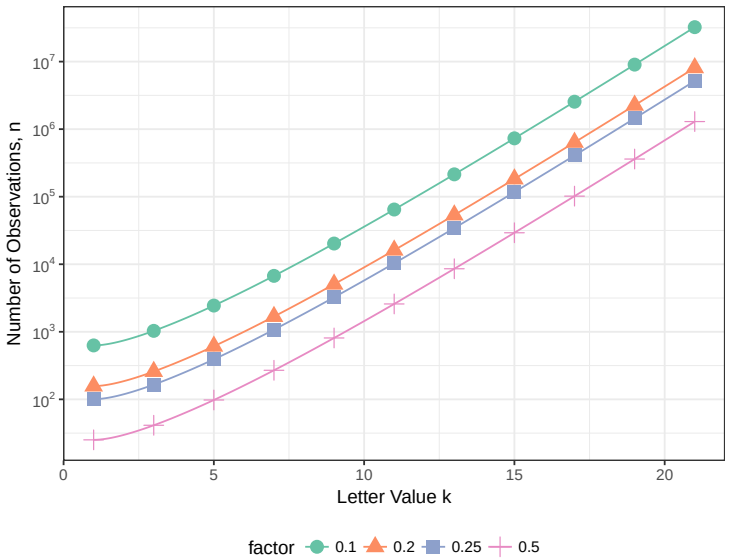



Figure 4: Plot of letter value vs. number of observations (on log-scale) needed for a 2-SD uncertainty of no more than 0.50, 0.25, 0.20 and 0.10 $\sigma$.

The different rules provide the user with choices depending on the desired precision of the displayed letter values. Our implementation defaults to rule 3 based on a 95% confidence level.

## 5 Examples

### 5.1 Heavy-tailed distributions and populations

This first example highlights the behavior of letter-value plots in data and distributions with heavy tails. Figure 5 which shows three decreasingly heavy-tailed $t$ distributions with 2, 3, and 9 degrees of freedom. In the letter-value plots, it is easy to see the difference from a Gaussian distribution for both the $t_2$- and the $t_3$-distributions due to the peak around the median, which emphasizes the relatively light middle of the distribution compared to the long tails.

Figure 6 shows letter-value plots and boxplots for the 1980 populations and log pop-

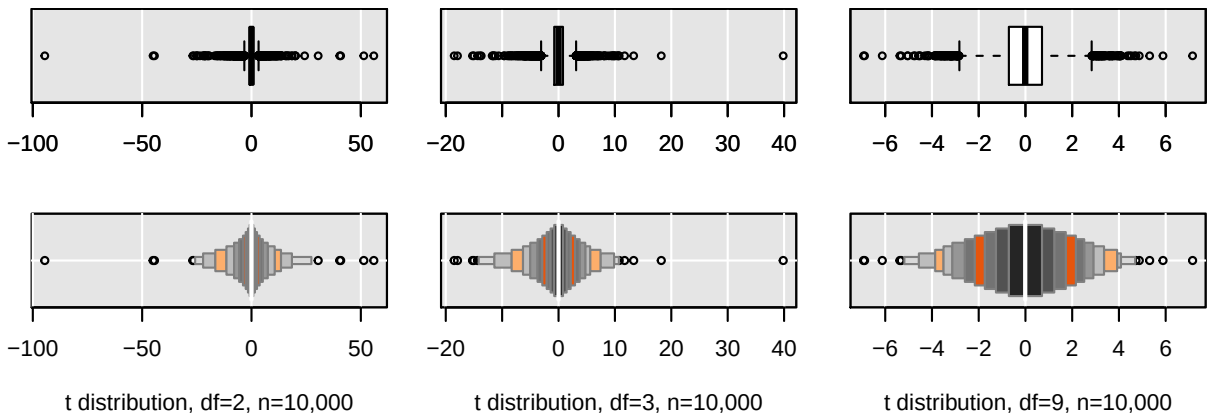



Figure 5: Letter-value plots and standard boxplots for samples of 10,000 for $t$ distributions on 2, 3, 9 degrees of freedom. Top row: Conventional boxplots. Bottom row: Letter-value plots.

ulations of the 3068 counties in the United States. While the skewness in the distribution of the populations is evident from both the letter-value plot and the conventional boxplot, the former shows more clearly that the right tail of the log populations above the median is somewhat more extended than the left tail below the median (i.e., the boxes to the right of the median are slightly longer than those to the left of the median).

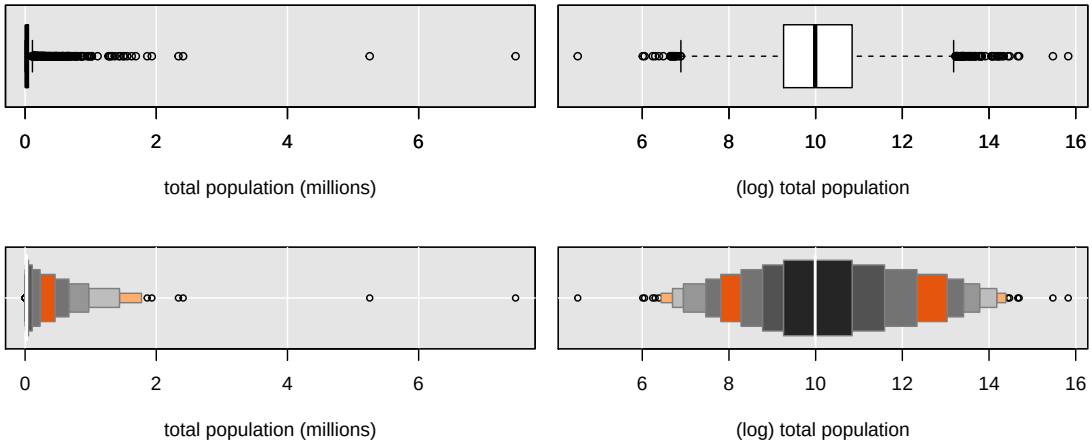



Figure 6: Letter-value plots and standard boxplots for the 1980 populations and log populations of 3068 counties in the continental United States.

### 5.2 Contaminated data

The second example demonstrates the behavior of letter-value plots in situations with contaminated/heterogeneous data. The plots in Figure 7 give an overview of the situa-

tion: starting with a sample of size 1,500 from a standard normal distribution, N$(0, 1)$, we replace values randomly by a set of points sampled from N$(4, 1)$. We start with 1% contamination (of the original sample size $n$) and end at 25%, which gives an essentially bi-modal distribution. Figures 8 show boxplots and letter-value plots for this data.

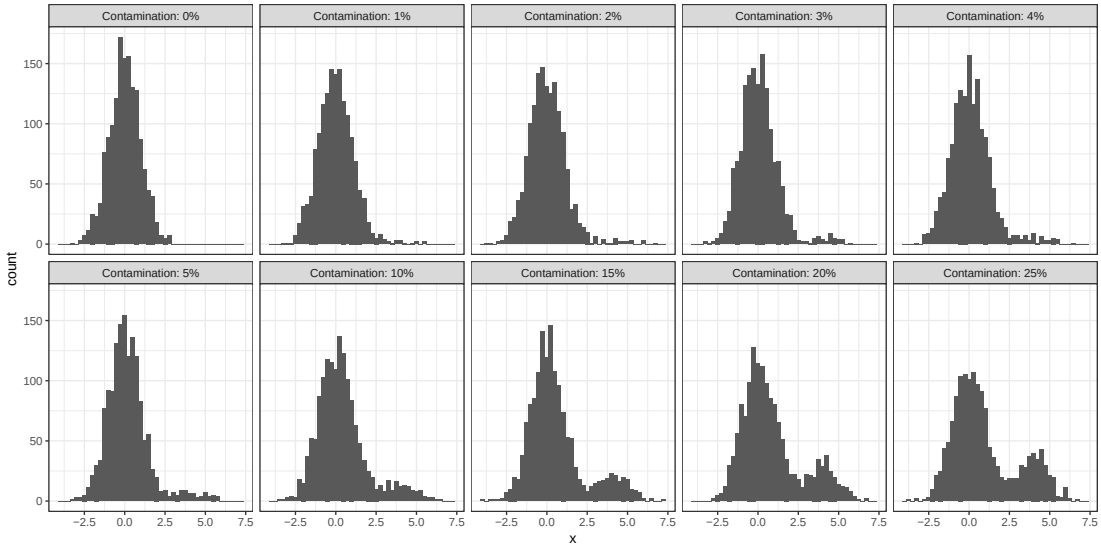
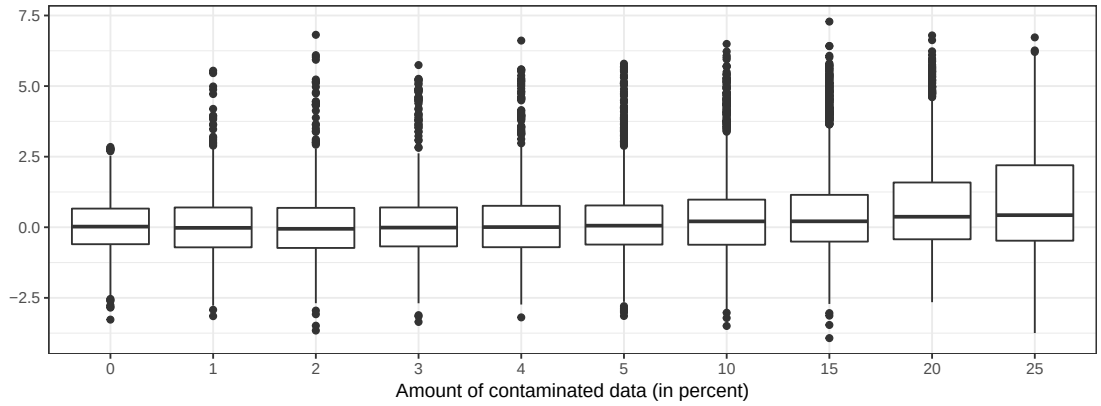


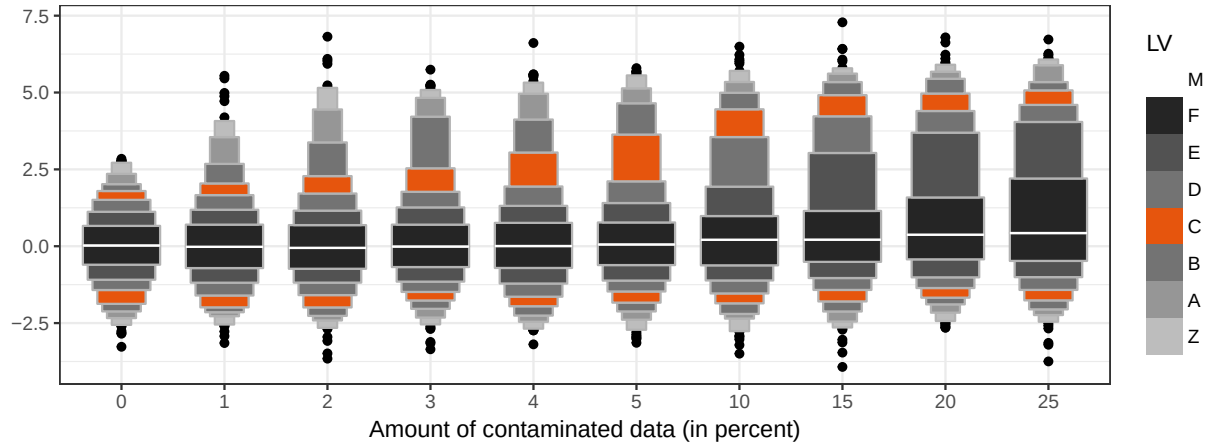

Figure 7: Histograms as an overview of the bi-modal samples resulting from contaminating a sample from a standard normal distribution with samples from $N(4, 1)$.

In the boxplots (see Figure 8a), we see that up to 5% of contamination in the upper tail area goes unnoticed in the median and the fourths of the boxplots. Up to 20% of contamination raises the median slightly and increases the inter-quartile range some at the cost of some of the outliers. At 25% the number of outliers becomes much smaller, while the upper whisker is extended to beyond the second mode and the upper quartile is raised. The letter-value plots of Figure 8b are more sensitive, and thereby are able to give a better summary of the actual underlying distribution. With 1% of contamination (i.e. 15 additional points from N$(4, 1)$), we see that the three outmost letter values are shifted upward – indicating an asymmetry in the data, that is restricted to the outer tail area. As the percentage of contaminated points gets higher, letter values at a higher depth are affected. For 25% contamination, we see that the upper eighth is shifted into the mean of the second mode. In the area-adjusted letter-value plots of Figure 8c the second mode in the data becomes visible for a 20% contamination in an increase of the widths of boxes associated with the letter-values of the upper tail.

(a) Boxplots of contaminated data. Boxplots are robust towards contamination of less than 25%.




(b) Letter-value plots with increasing amount of contamination.




(c) Area-adjusted Letter-value plots.

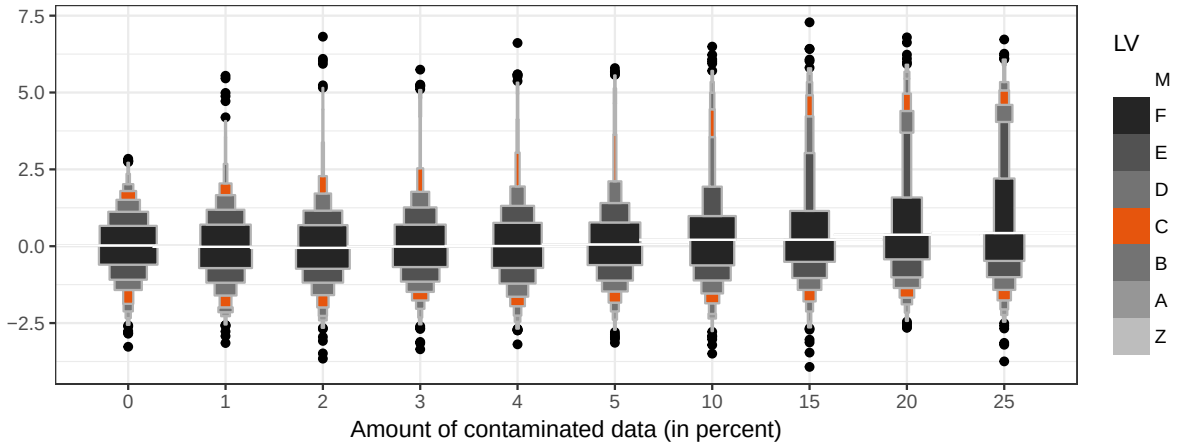



Figure 8: Boxplots and Letter-value plots of contaminated data.

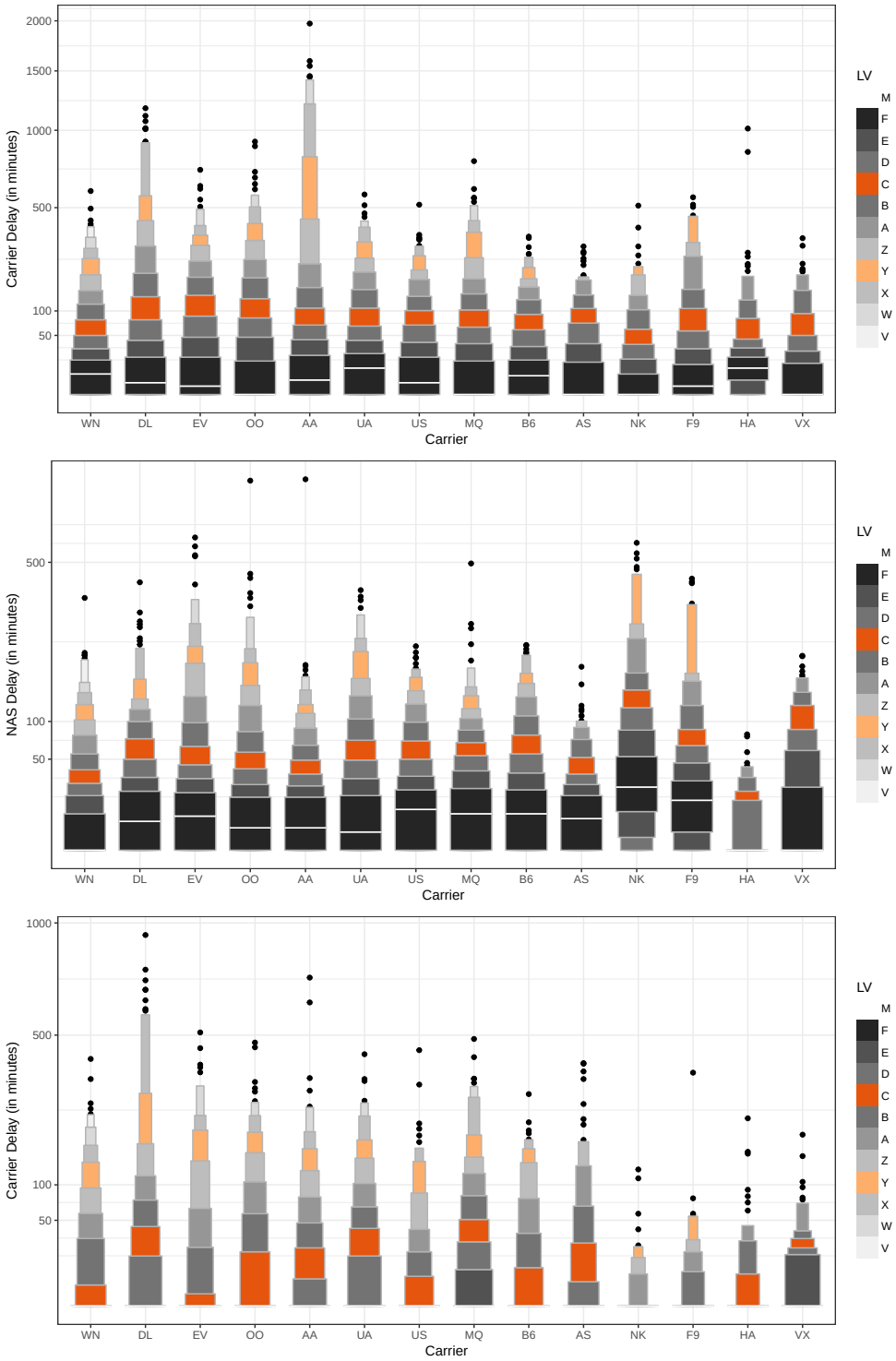



Figure 9: Letter value plots of different types of delays by airline carriers: carrier delay (top), National Air System (NAS) delays (middle), and weather related delays (bottom).

### 5.3 Zero-inflated data

The third example shows again an aspect of flight data: Figure 9 displays all 109,279 delayed flights in January 2015. As soon as a flight is delayed, FAA guidelines require airline carriers to report a reason for the delay. The three most common reasons for delays are inclement weather, carrier delays, and National Air System (NAS). While the different sources of delay have to be stated, delays of less than 15 minutes can be reported as zero, leading to highly zero-inflated distributions for these types of delays. Zero-inflations lead to letter-value plots that are seemingly 'cut off' – indicating that a lot of the lower letter values coincide.

From the letter-value plots in the top row of Figure 9 we see that most airline carriers have median delay times due to carrier delays close to zero, with the notable exception of Hawaiian Airlines (HA) and United Airlines (UA), where the median delay is close to $25 = 5^2$ minutes. HA is one of the smallest carriers, yet shows some of the most extreme delays, indicating a potential organizational problem. With the exception of Spirit Airlines (NK), all airline carriers seem to be affected similarly by delays due to National Air System (NAS), as we can see from the letter-value plots in the middle of Figure 9. There are some exceptions: Hawaiian Airlines has almost no problems due to NAS delays, Southwest (WN) and Virgin America (VX) have a median of zero delay due to NAS. The small carrier Spirit Airlines (NK) seems to be affected by NAS delays surprisingly often and long. Weather delays, as shown in the bottom row of Figure 9 affect different airline carriers differently, most likely due to different routes serviced. Hawaiian Airlines (HA) is not likely to be affected by inclement weather in January, and therefore delays due to weather are minimal. Envoy Air (MQ, formerly American Eagle Airlines) and Virgin America (VX) on the other hand, are affected the most by weather delays, and United Airlines (UA) has also a particularly long tail, indicating that some of its flights are affected very severely by weather events.

## 6 Conclusions

Letter-value plots provide a natural extension of boxplots in situations where we are dealing with large amounts of data. Like boxplots, they show only actual data values, rather than smoothed values or estimated densities. Letter-value plots convey more information about the tail, beyond the whiskers. Simple stopping rules that depend on neither the number of points nor on their distribution allow us to construct reliable plots that are less prone to over-interpretation when dealing with small number of points: stopping rule 3 ensures that a box for quartiles is drawn only if there are at least 16 data points. This rule is sensible in situations where we are dealing with groups of very different sizes. Additionally, for large data situations, fewer observations will be labeled as as "outliers" compared to a conventional boxplot, where there is a fixed rate of outliers – for a normal distribution it is approximately 0.7%.

Using location depths, introduced by Tukey (1975), letter values could be extended to bivariate situations, which would permit the inclusion of letter-values in bagplots (Rousseeuw et al., 1999; Wolf and Bielefeld, 2010).

### SUPPLEMENTARY MATERIAL

**R-package lvplot:** R-package lvplot consists of the implementation of letter-value plots as described in the paper. All data sets used for this paper are also included in the package. Version 0.2.0 of the package was used for all letter-value plots in the paper and is available from CRAN. (GNU zipped tar file)

**lettervalue.R:** R code to reproduce all of the figures in the paper. (.R file)

## References

Benjamini, Y. (1988), "Opening the box of a boxplot." *The American Statistician*, 42, 257–262.

David, H. A. and Nagaraja, H. N. (2003), *Order Statistics*, New York: Wiley Series in Probability and Statistics.

Emerson, J. D. and Strenio, J. (1983), *Boxplots and batch comparison*, chap. 3, in Hoaglin et al. (1983), pp. 58–96.

Esty, W. W. and Banfield, J. D. (2003), "The Box-percentile Plot," *Journal of Statistical Software*, 8.

Hintze, J. L. and Nelson, R. D. (1998), "Violin plots: A box plot–density trace synergism," *The American Statistician*, 52, 181–184.

Hoaglin, D. C. (1983), *Letter values: A set of selected order statistics*, chap. 2, in Hoaglin et al. (1983), pp. 33–57.

Hoaglin, D. C. and Iglewicz, B. (1987), "Fine-tuning some resistant rules for outlier labeling," *Journal of the American Statistical Association*, 82, 1147–1149.

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.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (eds.) (1983), *Understanding Robust and Exploratory Data Analysis*, New York: Wiley.

Hyndman, R. J. (1996), "Computing and Graphing Highest Density Regions," *The American Statistician*, 50, 120–126.

Kampstra, P. (2008), "Beanplot: A Boxplot Alternative for Visual Comparison of Distributions," *Journal of Statistical Software, Code Snippets*, 28, 1–9.

Mosteller, F. (1946), "On Some Useful 'Inefficient' Statistics," *Annals of Mathematical Statistics*, 17, 377–408.

Rousseeuw, P. J., Ruts, I., and Tukey, J. W. (1999), "The Bagplot: A Bivariate Boxplot," *The American Statistician*, 53, 382–387.

Tukey, J. W. (1960), *A survey of sampling from contaminated distributions.*, Stanford University Press, pp. 448–485.

— (1970), *Exploratory Data Analysis*, Addison–Wesley, preliminary ed.

— (1972), "Some Graphic and Semigraphic Displays," in *Statistical Papers in Honor of George W Snedecor*, ed. Bancroft, T. A., Ames, Iowa: The Iowa State University Press, pp. 293–316.

— (1975), "Mathematics and the Picturing of Data," in *Proceedings of the 1974 International Congress of Mathematicians*, ed. James, R. D., Vancouver, vol. 2, pp. 523–531.

— (1977), *Exploratory Data Analysis.*, Addison-Wesley.

Wickham, H. and Stryjewski, L. (2012), "40 years of boxplots," Tech. rep., had.co.nz.

Wolf, P. and Bielefeld, U. (2010), *aplpack: Another PLot PACKage: stem.leaf, bagplot, faces, spin3R, and some slider functions*, R package version 1.2.3.