Letter-value plots: Boxplots for large data

Heike Hofmann
Dept of Statistics
Iowa State University
Ames, IA 50011

Karen Kafadar
Dept of Statistics
Indiana University
Bloomington, IN 47408

Hadley Wickham
Dept of Statistics
Rice University
Houston, TX 77001

December 2, 2011

Abstract

Conventional boxplots (Tukey, 1977) are useful displays for conveying rough information about the central 50% and the extent of data. For small-sized data sets ($n < 200$), detailed estimates of tail behavior beyond the quartiles may not be trustworthy, so the information provided by boxplots is appropriately somewhat vague beyond the quartiles, and the expected number of "outliers" of size n is often less than 10 (Hoaglin et al., 1986). Larger data sets ($n \approx 10,000$–$100,000$) afford more precise estimates of quantiles beyond the quartiles, but conventional boxplots do not show this information about the tails, and, in addition, show large numbers of extreme, but not unexpected, observations.

The letter-value plot addresses both these shortcomings: (1) it conveys more de-

tailed information in the tails using letter values, but only to the depths where the letter values are reliable estimates of their corresponding quantiles and (2) "outliers" are labeled as those observations beyond the most extreme letter value. All features shown on the letter-value plot are actual observations, thus remaining faithful to the principles that governed Tukey's original boxplot. We illustrate letter-value plots on real data (univariate and bivariate) that demonstrate their usefulness, particularly for large data sets. All graphics are created using R (R Development Core Team, 2011), and code and data are available in the supplementary materials.

Key words: boxplots, quantiles, letter value display, fourth, order statistics, tail area, location depth.

1 Introduction

Boxplots (Tukey, 1970, 1972) give a compact graphical summary of the distribution of a variable, based around a set of order statistics called letter values. In *Exploratory Data Analysis*, Tukey (1977) recommended the $([n/2]+1)/2$-th and $(n+1-([n/2]+1)/2)$-th order statistics as estimates of the quartiles, the "lower fourth" and "upper fourth" in Hoaglin et al. (1983) with *depth* $([n/2]+1)/2$ because they lie that many observations in from the extremes.

Boxplots 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 behaviour are not displayed, despite the fact that larger sample sizes allow more reliable estimates further out into the tails. Figure 1 illustrates both problems with a boxplot of 135,605 internet (log-transformed) session durations, stratified into 32 groups based on the logarithm of the number of bytes transferred during the session. See Kafadar and Wegman (2004) for further details about the data and transformations. The sample sizes in the 32 boxes range from 1341 (box #32) to 7865 (Box #13), with a median sample size of 4238. With so many observations in each category, the number of labeled outliers is huge, with far too many to investigate individually, making it difficult to distinguish between extreme values and true outliers.

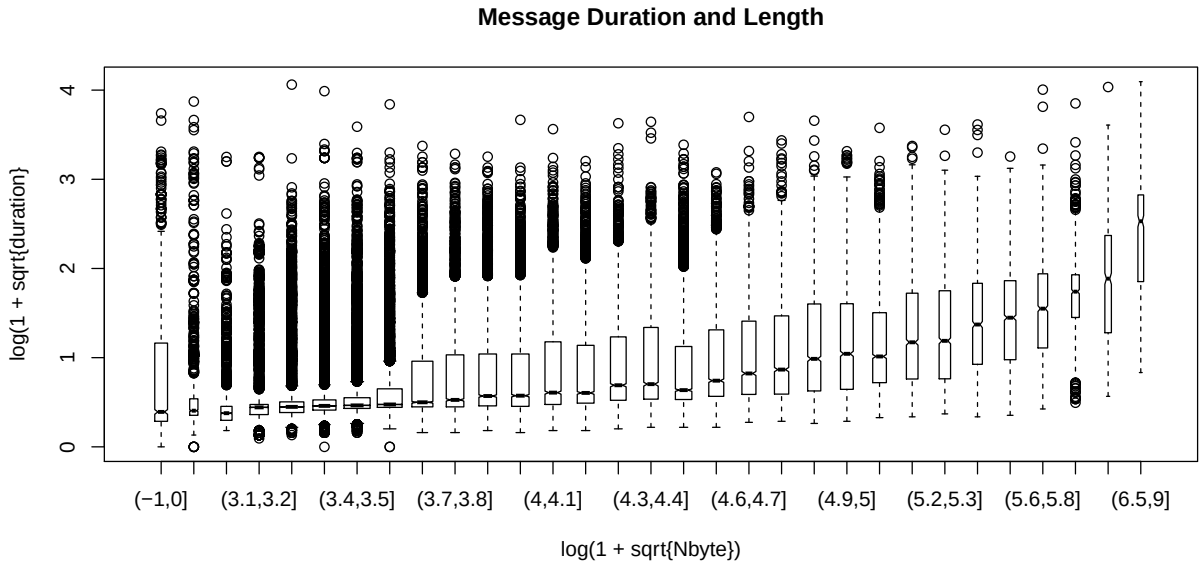

Figure 1: Notched boxplots (McGill et al., 1978) of (log-transformed) duration for 135,605 internet sessions, grouped by ranges of (log-transformed) byte lengths for the sessions.

An observation is labelled as an “outside value” (which we will denote here simply as

"outlier") 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 same the distribution as the bulk of the data; or (b) true outliers, 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 "outliers" grows approximately linearly with n: the theoretical fourths from a sample of independent Gaussian observations are $\pm 0.6745\sigma$, yielding an interquartile range of 1.35σ, and inner fences at $\pm(0.675 + 1.5 \cdot 1.35)\sigma = \pm 2.70\sigma$. Therefore the box and whiskers covers 99.3% of the distribution, leaving about 0.7% of the points to be labeled as "outliers" (cf. Hoaglin (1983)). Similarly, the probability of getting at least one "outlier" for Gaussian data exceeds 30% for samples of size 50, and 97% for samples of size 500 (Hoaglin et al., 1986, pg. 1148). The approach of Hoaglin and Iglewicz (1987), which labels a fixed number of "outliers" ("fixed outside rate") using a rule based on the fourths, avoids this dependence of the expected outside rate on the underlying distribution. Although it avoids the linear dependence of number of "outliers" on n, it 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 better illustrate tail behavior, such as vase (Benjamini, 1988), violin (Hintze and Nelson, 1998), and box-percentile plots (Esty and

Banfield, 2003). These displays provide more detailed information about the distributions, through the use of nonparametric density estimates, which is 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 with a variable number of letter values, selected based on the uncertainty associated with each estimate and hence on the number of observations. Any values outside the most extreme letter value are displayed individually. These two 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 the same data as a letter-value plot, which better shows the skewed tails (even with the logarithmic transformations) and far fewer "labeled" outliers. Letter-value plots are described in 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 the successive letter values. Section 4 discusses three other rules to select the

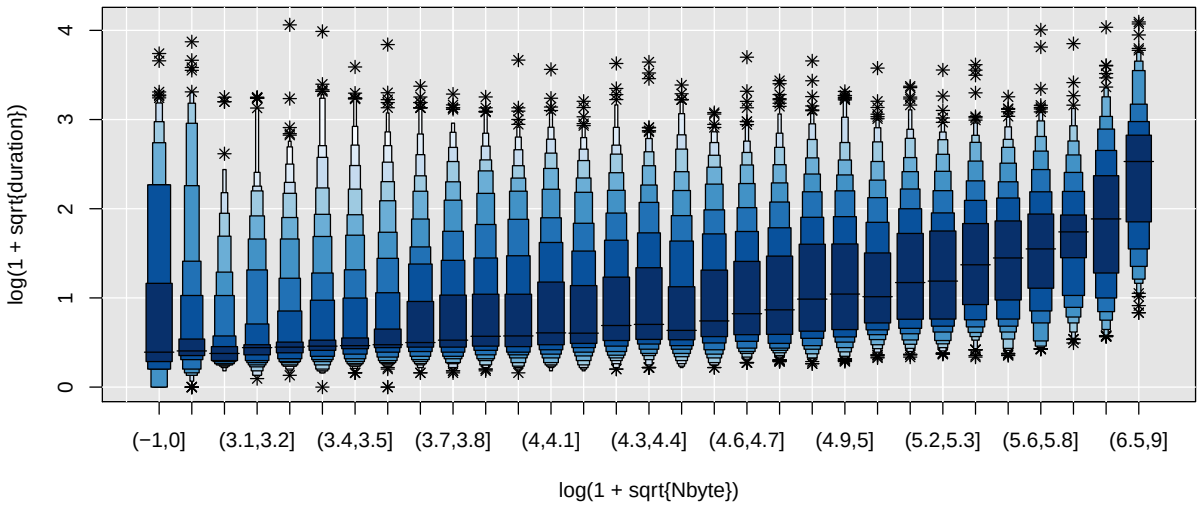

Figure 2: Letter-value plots of (log-transformed) duration for 135,605 internet sessions, grouped by ranges of (log-transformed) byte lengths for the sessions.

letter values based on the sample size. Some proposals for multivariate data and final discussion appears in Sections 5 and 6.

Our implementation of letter-value plots is available as an R package, `lvplot`, from CRAN. The online supplementary material contains all code and data used for the plots in this paper.

2 Letter values

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 greatest integer below y and the next integer above y, respectively. The letter values are those order statistics having specific depths, defined recursively starting with the median. The depth of the median, d_1, 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_1 and d_M denote the depth of the median.) If the depth is an integer plus $\frac{1}{2}$, then the lower letter value is defined as the average of the two adjacent order statistics, $X_{(\lfloor d_i \rfloor)}$ and $X_{(\lceil d_i \rceil)}$, 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 $U_i = X_{(n-d_i+1)}$. The advantage of this definition for the letter values is that the median of the sampling distribution for this sample quantile from a continuous distribution $F(\cdot)$ is very close to $F^{-1}((i - \frac{1}{3})/(n + \frac{1}{3})$, for a wide range of F, n, and i (Hoaglin, 1983). 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. Moreover, 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 $[(0.25 \cdot 0.75)/(n\phi(\Phi^{-1}(0.25)))]^{1/2}\sigma = 1.362\sigma/\sqrt{n}$ or a 2-SD uncertainty of

roughly 0.25σ 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 20 letter values, as well as the factor in 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 0.25σ; we would need a sample of size 1.4·120 = 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 about all the information that the data can reliably support.

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. Figure 3 illustrates this retention of tail information in visualizing the distribution of the 1980 populations and their logarithms in 3068 continental U.S. counties via normal quantile-quantile (QQ) plots (panels A and B, respectively) versus using only the 25 letter values (panels C and D, respectively); the

LV	ideal tail area	rough %	odds (2^i)	SEfactor	n-equiv*
M	.50	50.0%	2	1.253314	
F	.25	25.0%	4	1.36	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
Y	.001953125	0.2%	512	7.11	27.3
X	.0009765625	0.1%	1,024	9.48	48.4
W	.00048828125	0.05%	2,048	12.70	87.0
V	.000244140625	0.024%	4,096	17.11	157.7
U	.0001220703125	0.012%	8,192	23.14	288.5
T	.00006103515625	0.006%	16,384	31.40	531.3
S	.000030517578125	0.003%	32,768	42.75	984.4
R	.0000152587890625	0.0015%	65,536	58.34	1833.5
Q	.00000762939453125	0.0008%	131,072	79.80	3430.5
P	.000003814697265625	0.0004%	252,144	109.38	6444.3
O	.0000019073486328125	0.0002%	504,288	150.19	12149.2
N	.0000095367431640625	0.0001%	1,008,576	206.55	22977.6

Table 1: First 20 letter values. Ideal `tail area` is 2^{-i}, $i = 1, ..., 20$. `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 σ^2) corresponding to `tail area`, i.e., $SE(LV) \approx$ `SEfactor` $\times\sigma/\sqrt{n}$, where `SEfactor` = $\sqrt{p_i(1-p_i)}/\phi(\Phi^{-1}(p_i))$, p_i = `tail area` = 2^{-i}. `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; e.g., need 1.4n (respectively, 2.1n) observations for the eighth (respectively, sixteenth) to have the same uncertainty as that of a fourth from a sample of size n.

right column reveals the advantage of logarithms.

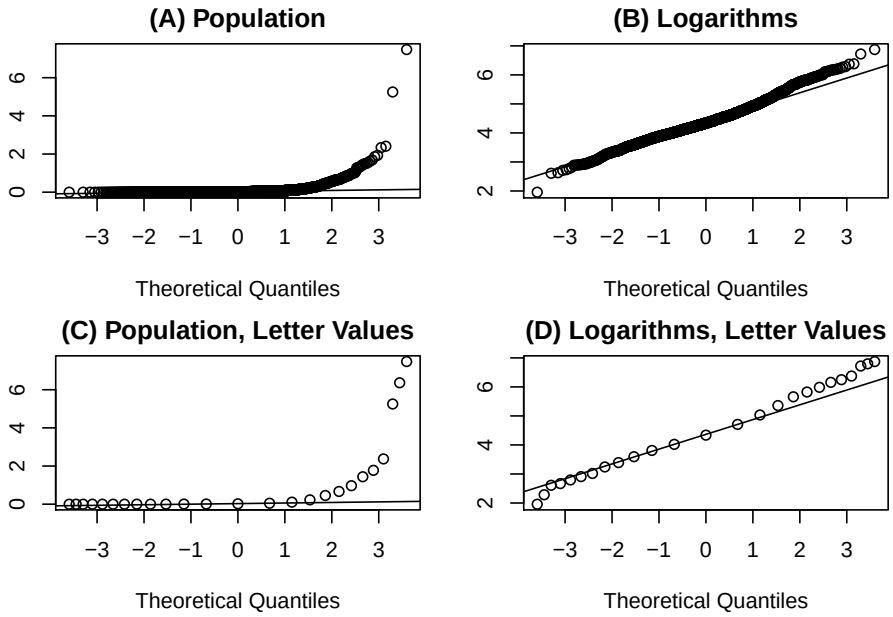

Figure 3: QQ plots on 3068 Continental U.S. county populations (A) and their logarithms (B), versus QQ plots (C), (D) using only 25 letter values.

3 Letter-value plots

Letter-value plots are based on the letter values, with one box for each pair of lower and upper letter values. The median is shown by a vertical line segment, and the innermost box is drawn at the lower and upper fourths, as in the conventional boxplot. An incrementally narrower box is drawn between at the lower and upper eighths, and narrower one still at the lower and upper sixteenths. We continue in this fashion until we reach a box that corresponds to a stopping rule described in the following section. Boxes with matching heights correspond to the same depths. We also shade more heavily the innermost boxes, to indicate a higher data density.

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 d_i reaches 1, the letter values are the extremes (minimum and maximum).

Letter-value plots for three different distributions are shown in Figure 4. Each panel displays a sample of 10,000 data points (top = standard Gaussian, middle = exponential with mean 1, bottom = standard uniform), first using the proposed letter-value plot up to letter Y, corresponding roughly to tail area $2^{-9} = 1/512$ (top row), and then with the the conventional boxplot (bottom row). Comparing the left (Gaussian) and right (Uniform) letter-value plot, overall *more heavily shaded* displays correspond to distributions with *lighter* tails. This phenomenon is shown more forcefully in Figure 5 which shows three decreasingly heavy-tailed t distributions with 2, 3, and 9 degrees of freedom.

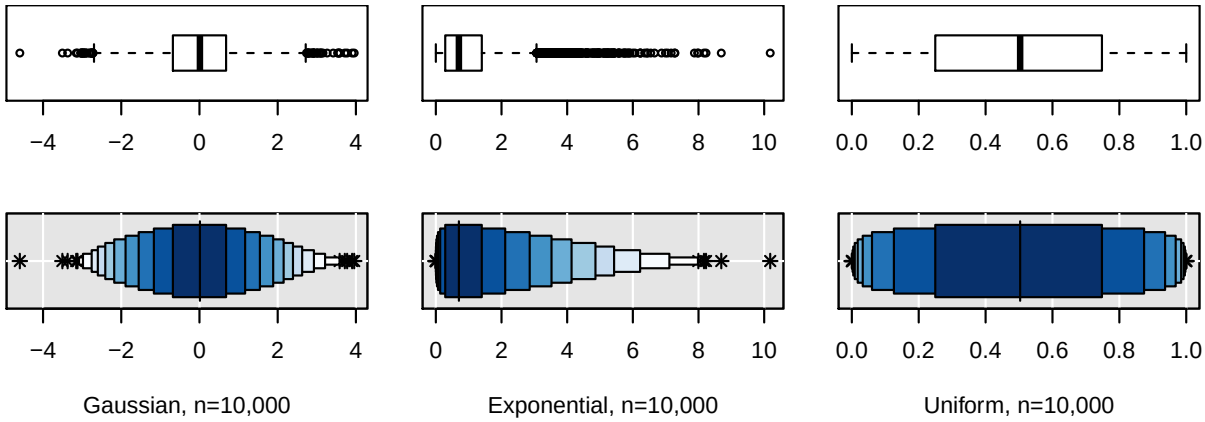

Figure 4: Letter-value plots (top row) and standard boxplots (bottom row) for data from three different distributions. Each plot shows 10,000 data points. From left to right, samples come from $\mathrm{N}(0, 1)$, $\mathrm{Exp}(1)$, and $\mathrm{U}[0, 1]$.

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

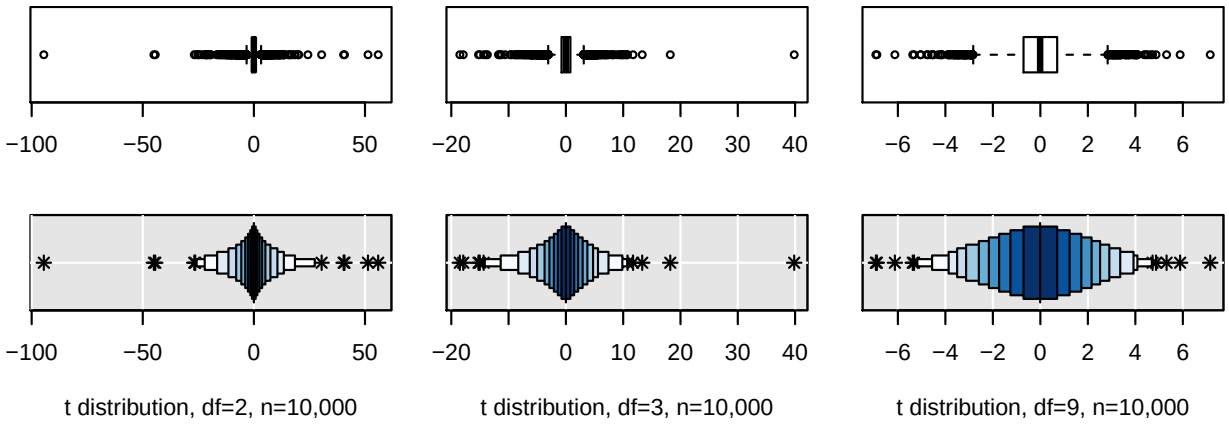

Figure 5: Letter-value plots and standard boxplots for samples of 10,000 for t distributions on 2, 3, 9 degrees of freedom. Top row: Letter-value plots. Bottom row: Conventional boxplots.

lations 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).

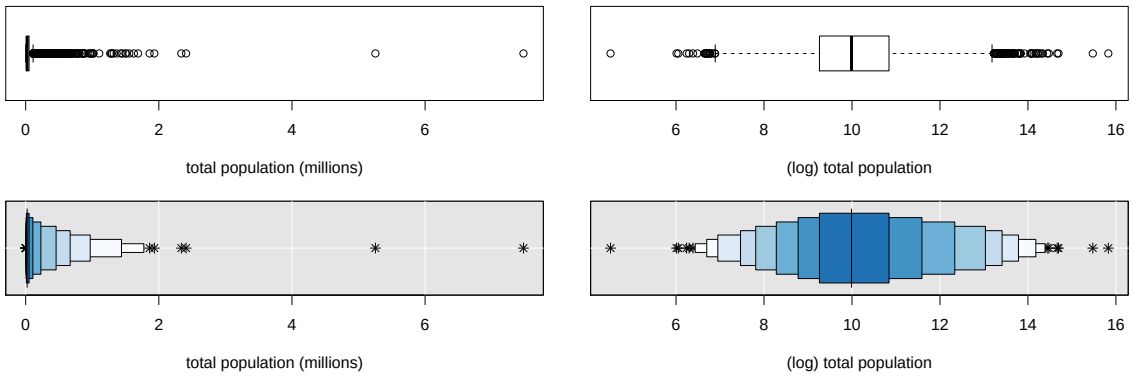

Figure 6: Letter-value plots and standard boxplots for the 1980 populations and log populations of 3068 counties in the continental United States.

4 Extent of letter-value plots

We need a rule to determine the number of boxes to show in a letter-value plot, which will determine the number of labeled outliers. In this section, we consider four proposals for such a rule.

In many of the displays in *Exploratory Data Analysis*, Tukey identified 5–8 extreme points. As a rough guideline, 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. Recall that the depth of the k^{th} letter value, d_k, is defined in terms of the previous depth: $d_k = (1 + \lfloor d_{k-1} \rfloor)/2$, which implies that $2d_k - 1 \leq d_{k-1} \leq 2d_k - (1/2)$. If we stop the letter-value plot display at LV_k where

$$k = \lfloor \log_2 n \rfloor - 3 \tag{1}$$

then we can expect to label 5–8 observations in each tail.

An alternative criterion fixes the number of labeled outliers as a percentage of the overall sample size. Letting p denote this proportion, the last set of boxes to be drawn ends with depths

$$k = \lfloor \log_2 n \rfloor - \lfloor \log_2(np) \rfloor + 1 \tag{2}$$

In effect, conventional boxplots use this rule, with $p = 0.007$. This criterion results in the same rule as the previous rule for the samples in Figure 4 ($n = 10,000$) when p lies

between 0.004 and 0.006 (0.4–0.6%), and in Figure 6 ($n = 3068$) when p lies between 0.011 and 0.020 (1.1–2.0%).

A third approach is a rule in which the final k is based on the "trustworthiness" of the k^{th} letter value as an estimate of the corresponding population quantile. "Trustworthiness" can be characterized by the approximate 95% confidence interval around a given letter value: if the interval overlaps the subsequent letter value, then the uncertainty for the given letter value is high enough that we should not display it. Thus, boxes are shown for those letter values whose approximate 95% confidence intervals exclude the neighboring letter values. Since a letter value can be viewed as the median between the extreme and the previous letter value, and since an approximate $1 - \alpha$ confidence interval for the median of m values (with $m > 10$) has approximately $0.5\sqrt{m}z_{1-\alpha/2}$ observations on both sides 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, 161), based on the Gaussian approximation to the binomial distribution), this third criterion leads to a particularly straightforward rule. (This result from David and Nagaraja (2003, 161) assumes a random sample, which is not true of the d_i observations beyond LV_i. A brief simulation confirmed that the result still holds sufficiently well for these purposes; details available from the authors.)

Consider the upper k^{th} letter value, LV_k. If its upper 95% confidence limit does not extend beyond the next letter value, LV_{k+1}, we continue to the next letter value, LV_{k+1}; otherwise, the box corresponding to LV_k is the last box shown. Since approximately d_{k+1}

observations lie between LV_k and LV_{k+1}, and the upper 95% limit for LV_k has roughly $\sqrt{d_{k-1}} \approx \sqrt{2d_k}$ observations, this principle requires $\sqrt{2d_k} < d_{k+1} \approx d_k/2$, or $d_k > 8$. A rule such as this one with 95% confidence level often leads to labeling 5–8 of the most extreme observations on each side, surprisingly consistent with many of the displays in Tukey (1977).

Generally, the third rule suggests showing k "trustworthy" LVs where $d_k > 2z^2_{1-\alpha/2}$, leading to the following stopping criterion:

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

This third stopping rule has the obvious advantage that it provides a simple, distribution-free solution. Neither the overall sample size, nor any distribution-related characteristic such as skewness or kurtosis, affects the rule. 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). Because the first rule is a special case of this third rule, we will consider the use of only stopping rules 2 and 3. Note that $k = 7$ when n is between 492 and 983, which corresponds to letter value A (Gaussian tail area 0.78%). When n is between 984 and 1966, $k = 8$, corresponding to letter value Z (Gaussian tail area 0.39%), which is very close to the expected percentage of outliers from a Gaussian sample that are labeled by the conventional boxplot rule (Gaussian tail area 0.35%).

A fourth rule is a variant of the previous rule, but assesses "trustworthiness" in terms of the standard error of the letter value. Table 1 shows `SEfactor`, the factor used for the

asymptotic standard error of the letter value for a Gaussian population:

$$SE(LV_i) \approx \sigma\sqrt{p_i \cdot (1-p_i)/n}/\phi(\Phi^{-1}(p_i)) = (SEfactor)\sigma/\sqrt{n}\,, \;\; p_i = 2^{-i}\,. \qquad (4)$$

When $i = 2$ ($p_i = 0.25$, fourths) and $n = 120$, this standard error is approximately 0.125σ; when $n = 186$, it is 0.10σ, and when $n = 743$, it is 0.05σ. Thus, a rough 2-SE interval around the fourth is roughly 0.25σ, 0.2σ, or 0.1σ, respectively, as n increases from 120 to 186 to 743. How many more letter values can be shown with the same level of uncertainty when n increases? For illustration, consider only the last (and most stringent) criterion, where $2SE \approx 0.1\sigma$. If $n \geq 1032$, the asymptotic 2-SE uncertainty in LV_3, the eighth (E), using the formula in (4), does not exceed 0.1σ. For the same precision in LV_4 (sixteenth), one needs $n \geq 1550$.

Table 2 lists the letter values and the ranges on n for which the uncertainty displayed up to a given letter value does not exceed 0.5σ, 0.25σ, 0.20σ, and 0.10σ. For example, when n = 10,000 and a 2-SD uncertainty around the letter value no greater than 0.20σ, one can show up to letter value 10 (X, 1/1024), but only up to letter value 7 (A, 1/128) for uncertainty of that does not exceed 0.10σ. Figure 7 plots the columns in Table 2 on a $\log_{10}$ scale, which shows that the logarithm of the sample size is approximately linear in the letter value number. This rule is very similar to rule (1) when the desired uncertainty in the letter values does not exceed 0.25σ.

The different rules provide the user with choices depending on the desired precision in the letter values shown. Our current implementation of the letter-value plot display uses

	i	0.5	.25	0.2	0.1
M	1	25	101	157	628
F	2	30	119	186	743
E	3	41	165	258	1,032
D	4	62	248	387	1,550
C	5	98	391	611	2,445
B	6	160	640	1,000	3,999
A	7	269	1,077	1,682	6,728
Z	8	463	1,851	2,893	11,570
Y	9	810	3,240	5,063	20,251
X	10	1,438	5,752	8,988	35,953
W	11	2,583	10,333	16,146	64,584
V	12	4,686	18,744	29,288	117,152
U	13	8,570	34,282	53,565	214,260
T	14	15,784	63,137	98,652	394,609
S	15	29,246	116,983	182,785	731,141
R	16	54,470	217,880	340,437	1,361,748
Q	17	101,914	407,654	636,960	2,547,840
P	18	191,449	765,796	1,196,557	4,786,227
O	19	360,931	1,443,726	2,255,822	9,023,287
N	20	682,624	2,730,498	4,266,403	17,065,610

Table 2: Letter values and sample size needed for 2-SE intervals of size 0.5 σ, 0.25 σ, 0.2 σ, and 0.1 σ, respectively

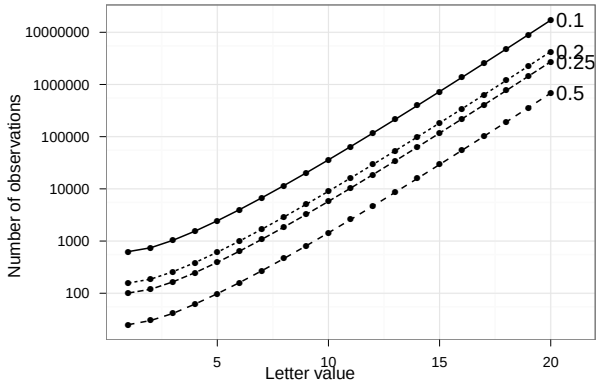

Figure 7: Plot of letter value vs. number of observations (on log-scale) needed for a 2-SD uncertainty of no more than 0.50, 0.25, 0.20 and 0.10 σ.

rule 3 as a default.

5 Bivariate letter value plots

Rousseeuw et al. (1999) proposed the "bagplot" as a two-dimensional version of the boxplot, using location depths (introduced by Tukey (1975)) to define analogues of the median and fourths, and then connecting the points corresponding to the fourth-depths via linear segments.

The location depth $ldepth(p, Z)$ is defined for an arbitrary point $p \in \mathbb{R}^2$, relative to a set of n points $Z = \{z_i = (x_i, y_i),\ i = 1, ..., n\}$, as the smallest number of z_i's contained in any closed halfplane with boundary line through p. That is, if one were to pass a line in the plane through p and keep track of the smaller of the two numbers of z_i on either side of the line as the line is rotated through every angle to its opposite side (180^o), then the location depth of that point p relative to Z is the smallest of all the numbers. The analog of the one-dimension median in two dimensions using location depth would thus be that point p_M for which $ldepth(p_M, Z)$ is largest (e.g., $n/2$). If such a p is not unique, then the "depth median" is defined as the "center of gravity" of all points p for which $ldepth(p, Z)$ is largest. Recall that a property of the fourths for a univariate sample is that the interval between the lower and upper fourth contains one-half of the data. Thus, an analog of the box for the bagplot was defined as the convex hull of all points p for which $ldepth(p, Z) \geq 1 + \lfloor 0.5 \cdot ldepth(p_M, Z) \rfloor$. Similarly, successive letter areas are defined based purely on their depth as the convex hull of all points in the sample with $ldepth_i(p, Z) \geq$

$1 + \lfloor 0.5 \cdot ldepth_{i-1}(p, Z) \rfloor$. Figure 8 shows a side-by side comparison of a standard bagplot (left, Rousseeuw et al. (1999), implemented in Wolf and Bielefeld (2010)) of average temperatures in January by degree latitude of each of 3,068 US counties and a letter-value bagplot on the right. The asymmetry of the bivariate display around the principal axis, and the cluster of points above the upper right edge of the final contour, are much more evident in the letter value bagplot. (These points correspond to ... ???) Algorithms for fast computation of bivariate letter values for a bivariate letter-value plot are currently under development.

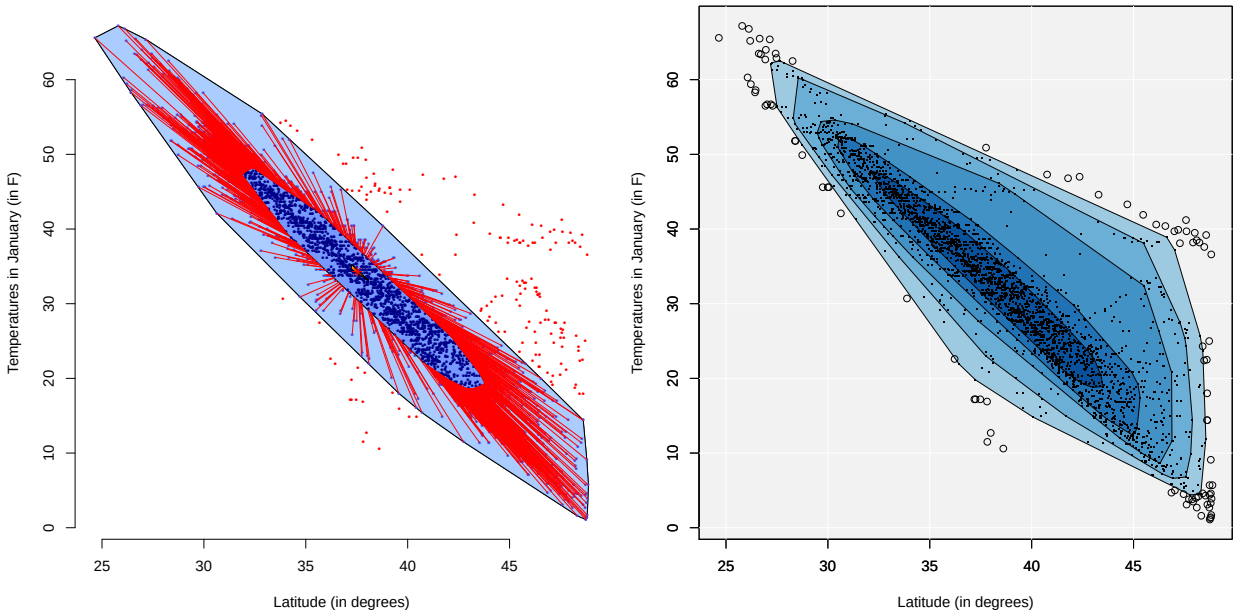

Figure 8: Bagplot (left) and letter-value bagplot (right) of average January temperatures by degree latitude for 3,068 US counties.

6 Summary

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 further information about tail behavior 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: Rule 3 will ensure 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, such as Figure 2. 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%. Letter values can be extended to two dimensions by using the location depth, giving rise to letter-value bagplots as a two dimensional extension of letter-value plots, and providing a robust, data-based assessment of data concentration in two dimensions. Implementation details are found in the online supplementary material.

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., 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.

Kafadar, K. and Wegman, E. J. (2004), "Graphical displays for internet traffic data." in *Proceedings of CompStat.*, pp. 158–170.

McGill, R., Tukey, J. W., and Larson, W. (1978), "Variations of box plots," *The American Statistician*, 32, 12–16.

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

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

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.

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.