Review of Manuscript TAS-11-231
Letter-value plots: Box plots for large data (datasets?)

This is an interesting idea whose potential usefulness is not fully demonstrated in this paper. The paper has a number of flaws that would need to be remedied before acceptance. Details follow:

1. Mark Twain described women's use of profanity, as "*they know the words, but not the music*." The same could be said of the author's use of EDA. EDA was conceived as a rough and ready way, in Tukey's words, "*To understand what we can do before we learn to measure how well we seem to have done it.*" Part and parcel of EDA is to see results and not have to read them. The development of more evocative plots fits well within this, but aspects of this development fly in the face of the underlying philosophy. First, do we really need big samples for EDA? Tukey's methods were aimed at pencil and paper – he continued doing this for many things long after he had high speed computing available. Not for all tasks, for he devoted an entire mainframe to spinning high dimensional data so he could look at three or even four-dimensional structure. But for most tasks we can sample from a large data set to better understand its structure. Indeed isn't one of statistics' primary goals to be able to make inferences about a lot from a little? I believe that the author(s) must demonstrate 2 or 3 real situations where looking at large datasets in toto is important, for it is only from these that we can learn important things.

2. I join with T. S. Eliot in believing that "*you gotta use words when you talk to me*." Before making definitions in symbolic form it is useful to explain, in words, what is going on, I understand that presenting just an equation is easier, but the goal of a paper is to make the reader's job easier, not the author's. And there are some terms that are introduced without defining them (e.g. whisker). I know that they are well-known in parts of the population, but I suspect that for those who know these words large parts of the paper are redundant. And so if you are writing just for them, at least 80% of the paper can be omitted. And even if everyone does know what a whisker is, there are babies being born every minute who do not. Define unusual terms.
3. *Caesar's wife must be beyond reproach*. If you are writing a paper on data display, your tables and graphics must be impeccable. The displays presented here are a long way from perfect. For example,

   Figure 1
   (i) The axis labels should not be in units – at most the units should be in parentheses – we'd never see an axis labeled, "Inches" it almost certainly would be "Height (in inches)". Here it should probably be "Information (in log of sqrt(bytes))".
   (ii) Second, the character of the distribution suggests that log is insufficient

(for the dependent variable) and should almost surely be inverse. If one does this the y-axis label could be "Speed" and be done with it.
(iii) And last, the example is contrived. Why use box plots at all when the x-axis is a continuous variable? A scatter plot seems more appropriate. Indeed, this might be a suitable example for the bivariate analogs you discuss in the last section. Find another example that is suitable for box plots.

Rounding – numbers presented for human eyes and minds should be rounded sensibly – surely not 0.6745σ; maybe 0.7 or, at most .67. Table 1 should be rounded to a level that readers would find useful – do we really need to know a tail area is .000000953674316406 25? Or is this just to show you have a sense of humor. Table 1 is almost certainly better as a figure. Also more evocative labels and names would be helpful I bet even the author can't remember what comes where. Do we need 20 levels?

Figure 2 Same problems as figure 1
Figure 3 Does the deviation from linearity in the extremes of the log QQ plots indicate that the letter values become less accurate in the tails?

4. Nits and typos
   Page 4 Hoaglin & Iglewicz ref is missing
   page 5 line 4 should read "…values selected on the basis if the uncertainty…"
   Page 6 9 lines up – notation for rounding up or down is too subtle. Good notation should be like a good teacher. This is too easy to miss. If you can't find some standard notation that does a better job, invent your own (append arrows?)
   Page 10 Opening sentences would do better at the very beginning of the article. You are burying the lead
   Page 11 Figures 4 and 5, in addition to the flaws specified earlier also has the caption reversed. It would be well to proof-read the manuscript more carefully before sending it in (note also the fragment on page 18 8 lines up).
   Should the title be "…*large data**sets***"?

5. Further suggestion
   I think there is a nice distinction to be made between an *outlier* (a truly rare event that occurs) and a *fringelier* (an unusual event that occurs more often than seldom). The former are easier to deal with, but the latter we really don't know what to make of them. Some graphic suggestions that emphasize the distinction would be useful.

So let me conclude:

1. This contains what seems like a good idea, but its usefulness needs to be demonstrated convincingly. It hasn't been yet. 2-3 good, real examples would help.
2. The writing needs to be less turgid.
3. The data displays need to be polished up and the analysis thought through more carefully.

When this is done I think there could be a fine paper that results.