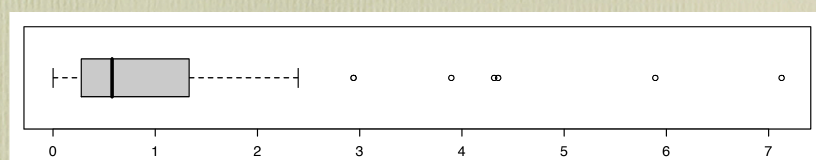
Letter Value Boxplot

Heike Hofmann, Karen Kafadar, Hadley Wickham
Iowa State University

Outline

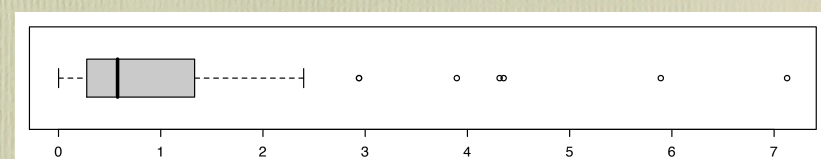
- Boxplots: Definition, Strengths & Weaknesses
- Letter Value Statistics
- Letter Value Boxplots
- Examples
- Conclusion

Boxplots



- Early Version: Tukey 1972 (Snedecor Festzeitschrift, at Iowa State University)
- Most common version in EDA (1977):
  - Median (Center Line), Fourths (Box Edges), adjacent values (ends of whiskers) and extreme values
  - All marks correspond to actual data values

Boxplot: Strengths

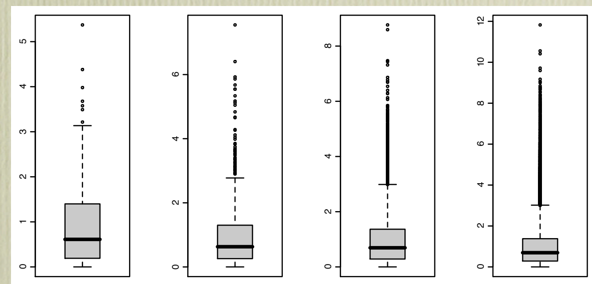


- Quick summary without overwhelming amount of detail
- Approximate location, spread, shape of distribution
- Outlier identification
- Associations among variables

Boxplots: Weaknesses

- Expected rate of labeled outliers approx 0.4+ 0.007n
- For n = 100000 expect approx. 700 outliers!

Exponential Distribution, n= 100, 1000, 10000, 100000



Modifications

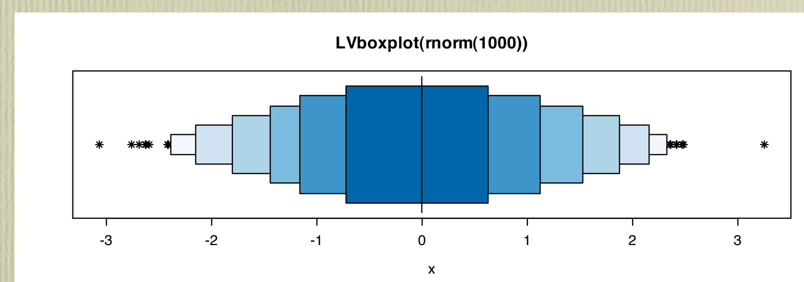
- Notched box-and-whisker (McGill, Larsen, Tukey 1987)
- Nonparametric density estimates
  - Vase plots (Benjamini, 1988)
  - Violin plots (Hintze, Nelson 1998)
  - Box-percentile plots (Esty, Banfield 2003)

Implementations: *S routines (David James), package* vioplot *(Adler, Romain), package* HMisc bpplot *(Harrell, Banfield), examples at R Graph Gallery*

Letter Value Statistics

- Estimate quantiles corresponding to tail areas 2^{-j}
  - Median (1/2): depth = $d_M = (1+n)/2$
  - Fourths (1/4): depth = $d_F = (1+\lfloor d_M \rfloor)/2$
  - Eights (1/8): depth = $d_E = (1+\lfloor d_F \rfloor)/2$
- Boxplots show median, fourths
- Large Data Sets: tail quantiles become more reliable $\longrightarrow$ include LVs beyond Fourths

Letter Value Boxplot



- How many boxes to show?
- Outlier identification?
- All marks are based on actual data values

Stopping Rules & Outliers

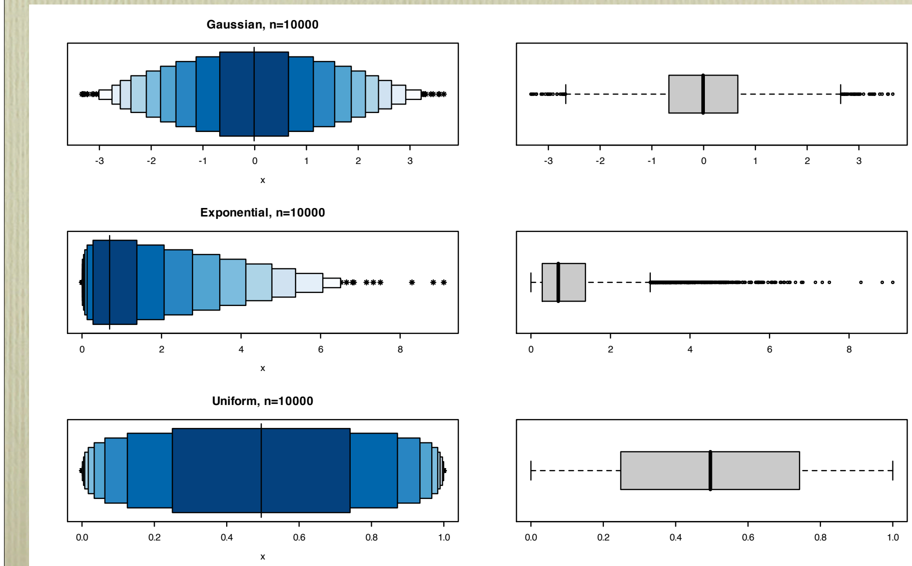
- EDA: 5-8 outliers $\longrightarrow$ $k = \lfloor \log_2 n \rfloor - 4$
- Percentage of data, e.g. 0.5-1%
- uncertainty in LV_i extends beyond or into LV_{i-1} (i.e. upper limit for LV_i crosses LV_{i-1})

$$\longrightarrow k = \left\lceil \log_2 n - \log_2 \left(4 z^2_{1-\alpha/2} \right) \right\rceil + 1$$

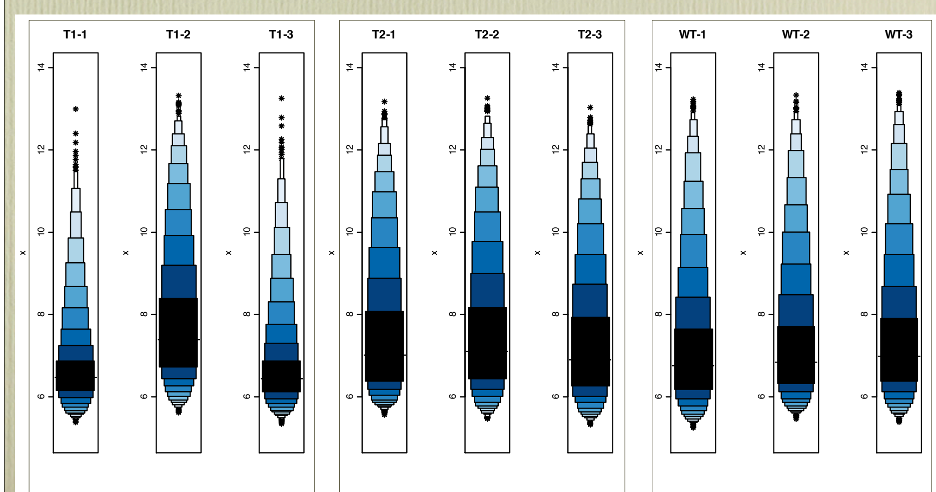
Rules lead to similar answers

... Examples

Gaussian, Exponential & Normal



Gene Expression Values



Conclusion

Letter Value Boxplots are

- appropriate for large number of values
- based on actual data values
- simple to compute
- reduce number of labeled outliers shown in conventional boxplots
- do not depend on a smoothing parameter

Download (for now) at http://www.public.iastate.edu/~hofmann

Graphical Displays of Large Data Sets

"The greatest value of a picture us when it forces *us to notice what we never expected to see"* (Tukey 1977)

- Quick summary without overwhelming amount of detail
- Approximate location, spread, shape of distribution
- Outlier identification
- Associations among variables