The second risk is server failure. If a strong server model is established so that client systems will not function without the server, and the server is not available, department members will sit around unable to use the system — just as we did twenty years ago when the mainframe was down. Many of us have memories of those days and no desire to return to them. The risk of server downtime is substantial. A good system administrator is the best protection against server downtime, since most server problems are related to software and its configuration —- not to hardware failure. One can buy a hardware maintenance contract and/or have spare parts available to recover from server hardware failure. Recovering from poorly configured server software is far more difficult.

Multi-platform issues

In any consideration of services being offered centrally from a department facility, thought should go into the variety of machines in the department. Will everyone benefit from a department mail server? If the server is a POP (Post Office Protocol) server, client software can be obtained for Mac, PC and UNIX machines, enabling all to take advantage of the new department service. Proprietary vendor email systems rarely cover all platforms.

On balance

The benefits of server installation outweigh the risks where qualified system administration can be dedicated to the task of maintaining the server. Strong server models are common because of the productivity advantages they deliver to their departments. Where system administration is weak or non-existent, the benefits of department servers are rarely realized.

Michael Conlon
Department of Statistics
Box 100212 HSC
University of Florida
Gainesville, FL 32610–0212
mconlon@stat.ufl.edu
Home page:
http://www.clas.ufl.edu/~mconlon

TOPICS IN SCIENTIFIC VISUALIZATION

A Colorful Variation On Box Plots

by Daniel B. Carr

Introduction and Background

The box plot provides a useful caricature for a batch of data. The word caricature is appropriate since the box plot hides some distributional features, such as gaps, while calling attention to others, such as the median and quartiles. Since box plots receive such widespread use, it is natural for statistical graphics designers to see if they can make even minor improvements on the graphical representation. This article describes my efforts to improve on the box plot. The article builds on the last newsletter article about grids (Carr 1994a). Both articles come from a technical report on converting tables into row-labeled plots. Interested readers can find more details and coverage of other topics such as representation of confidence bounds in Carr (1994b).

In terms of a brief history, Tukey (1977) introduced the box plot. McGill, Tukey, and Larsen (1978) added the idea of notches for comparing medians. Frigge, Hoaglin, and Iglewicz (1989) described different choices for the five- number summary portion of the box plot. Tufte (1983) provided alternative representations that reduced the amount of redundant ink in the box plot. His example suggested that others also search for a better graphical design. Tukey (1989) took up the challenge to develop a better representation. His design strategy used more easily distinguished graphical elements for different components of the box plot. In one variation an area symbol, the filled dot, represented the median. This area symbol stood out from other graphics elements and helped when comparing medians. Tukey (1993) provided more variations on letter displays and box plots and more thoughts on graphical comparison of several linked symbols. As discussants of Tukey's article, Becker and Cleveland (1993) reinforced the use of filled dots to represent medians and brought out further design considerations, such as sorting box plots by their medians. A simple look at the Becker and Cleveland example shows that filled dots and sorting provide a better basis for comparison. This history on black and white methods, the availability of color and variations that appear in software suggest that perhaps there is still room for improvement.

Design Objectives For The Box Plot

A useful starting point in graphical design is to review the design objectives. The box plot symbol is a composite symbol that represents five summary numbers (median, quartiles and adjacent values) plus outliers. The design of the box plot involves selecting a composite symbol to represent these five numbers. One design objective is for viewers to perceive the box plot as a single perceptual unit, as one symbol rather than five. A second objective is to ease comparison of elements within the box plot. For example, are the quartiles symmetric about the median? Other objectives involve comparisons among box plots: comparison of medians with other medians, upper quartiles with upper quartiles and so on. The multiple objectives complicate settling on an optimal design. Design efforts here focus on promoting comparisons among box plots.

Linkages that tightly bind elemental symbols into perceptual units make the elemental symbols difficult to disaggregate for comparison across perceptual units. Kosslyn (1994) discusses several principles of perceptual organization of which two, proximity and good continuation, are particularly relevant to composite symbol binding. For example, the box in the box plot binds the quartiles together. The visual flow along the box length and along the lines from the box to the adjacent values helps people perceive the outliers as part of the same distribution. A related principle for perceptual grouping is visual enclosure. Enclosing the median symbol in a box helps make the symbol become an integral part of the composite symbol. Weakening the tight binding provides a means of encouraging comparisons among box plots.

One design objective is for viewers to perceive the box plot as a single perceptual unit, as one symbol rather than five.

The filled dot suggested by Tukey (1989, 1993) is an area symbol that creates an alternative perceptual grouping:filled area symbols versus line symbols. The alternative grouping counteracts the tight binding in early versions of the box plot. This approach suggests the next step, using color to create alternative perceptual groups. When color is available, using an area symbol for the median becomes less advantageous. The box plot design shown in Figure 1 uses black vertical lines to call out medians. The design represents the ranges between the quartiles and adjacent values using filled boxes. Different colors distinguish the median, values below the median and values above the median. A simple white line shows the median comparison interval. The availability of color provides such options.

Several considerations motivate the particular design choices. Comparing medians across box plots is one of the most important comparison tasks. The current design gives the medians a different color and shape from the other box plot elements. The distinctive symbol enables people to rapidly find the medians using preattentive vision (see Julesz 1986). One can focus attention on all the black line segments just as one could focus attention on all filled dots. In contrast to the early box plot design, the line for the median comprises the vertical extremes of the composite symbol. This avoids total entrapment of the median symbol within the composite symbol. The vertical line segment orientation encourages the vertical visual flow needed to make comparisons. Note that by itself the horizontal orientation of the composite symbol impedes vertical visual flow. When comparing medians from adjacent rows the separation of the nearest endpoints is nearly horizontal. This helps in judging the correct distance. In the dot design it is difficult to judge the distance between dot centers. When adjacent medians are nearly identical in value, they can be resolved with vernier resolution. Thus the symbol for the median facilitates comparison.

As for the thicker box between the quartiles than between the adjacent values, Tukey (1993) points out that symbols near the edges of the plot have higher visual emphasis. Stressing symbols in the center of the distribution provides balance. Turning the lines from quartiles to adjacent values into thin boxes is a small change that accommodates the drawing of median comparison lines as described below.

Distinguishing the left and right sides of the box plot with color makes it easier to make comparisons among the respective halves of the distributions. For tasks like comparing third quartiles, one restricts attention to light red symbols. The almost instantaneous sorting by color narrows the field for comparisons. Evaluation of the interquartile range for a single box plot remains straight forward since the vertical line for the median and the color change does not seriously interrupt the visual flow along the thick box between the quartiles. A few considerations should govern the color selection. The colors should be easy to distinguish from each other and the median color. Colors for low and upper values should be brightness matched (see Carr 1994c) and not fully saturated (see Tufte 1990). When gray is the only option, the most important objective is make the median a distinctive color. This suggests dropping the color distinction between upper and lower values in gray level plots.

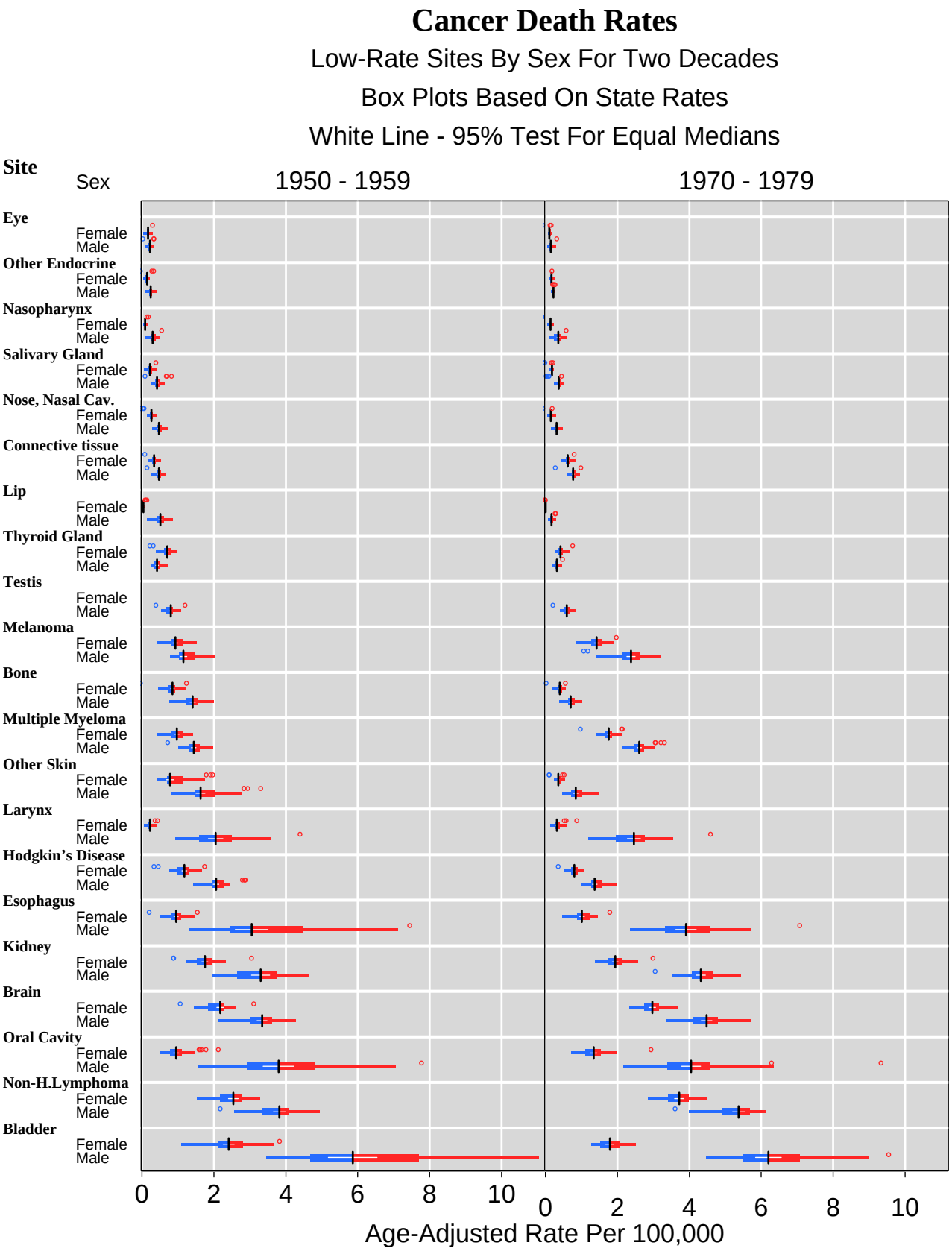

Figure 1: A Three-Factor Plot Showing Box Plots

The treatment of outliers in Figure 1 is standard. An alternative shows outliers as vertical line segments with the same height as the adjacent value box. This reduces overplotting, encourages vertical visual flow and plots symbols at their location centers rather than around them. Open circles help increase visibility when the segments are very short.

The proposed box plot design needs to be tested in cognitive experiments to see if it helps people to compare box plots more quickly and accurately.

Figure 1 uses a white horizontal line at the vertical center of each box to represent the median comparison interval. This line representation works when a confidence bound goes past a quartile. Notches appear strange when they extend past quartiles. The internal line representation has merits in addition to handling asymmetric distributions. Since the confidence limits are symmetric about the median, the confidence line allows more accurate comparison of symmetry for the first and third quartiles. To assess quartile symmetry, one looks at the difference between the ends of the white line and the ends of the quartile box. According to Weber's law, people can compare these small differences more accurately (on an absolute scale) than they can compare the larger distances to the median. The symmetry of the white line about the median encourages mentally folding the box plot about the median to judge symmetry. Since the white line is trapped inside the composite symbol, it detracts little when comparing symbol outline elements such as third quartiles. With median comparison lines inside the boxes, comparison of lines may not be as precise or easy as with other possible designs. This is a reasonable compromise because approximate hypothesis tests are not the primary box plot feature. In other words the comparison lines are easy to ignore but serve several comparison purposes when needed.

The interpretation of median comparison lines is the same as for notched box plots. The medians differ at the given significance level when the lines do not overlap and adjustments should be made in multiple comparison situations.

The Data Set, Plot Construction and Interpretation

The data set includes age-adjusted cancer rates for four factors: states (50), body sites (34), sexes, and decades (3). Figure 1 represents states using box plots, body sites and sex using indented rows and decades using panels. The 68 body site by sex combinations stretch the limits of the legible labeling and symbol plotting. Consequently this plot shows values for the 21 body sites with the lowest rates. Sorting by the male median values for the 1950's determines the body site order. The range of death rates remains large for these sites and a log scale would be reasonable for the x-axis. Figure 1 does not use the log scale in order to show the symmetry of the confidence lines and the performance of some box plot symbols under restricted resolution conditions. Figure 1 shows just two decades to provide more resolution for some symbols. As it stands the one page plot represents over forty percent of the 10,200 values in the data set.

The interpretation is straight forward even in the overplotting cases. For cancer of the eye, the median lines completely cover the quartiles and the median comparison lines. For males and females one can see the horizontal difference between the median lines. This indicates that the hidden comparison lines do not overlap. The conclusion is that rates for males and females are significantly different. For the twenty comparable body sites in the plot, male rates are higher with one exception, the thyroid gland. The plot design makes it easy to compare the rates for sexes. Comparing body site rates for a given decade and sex is not much harder since all values have the same scale. Comparisons across panels is a little harder but here the grid lines come into play (see Cleveland 1993 and Carr 1994a). Comparing box plot elements against the grid lines makes the rate increases from the fifties to seventies quite apparent.

Figure 1 provides a tremendous amount of information but is not a stopping point. To go further, one should study the national rates directly. The box plot outliers represent atypical state rates, so are of some interest. However careful comparison adjusts for estimate variability that includes the influence of differing population sizes.

Closing Comments

The plot that cannot be improved is rare. At the same time attempted improvement do not always prove successful. The proposed box plot design needs to be tested in cognitive experiments to see if it helps people to compare box plots more quickly and accurately. In the mean time, readers who like the design should give it a try. S-Plus users can modify the template for Figure 1. The cancer data, script file, rowplot function and box plot function are available by anonymous ftp to galaxy.gmu.edu. The directory `/pub/submissions/rowplot` contains this and many other row-labeled plot examples.

References

Becker, R.A. and W.S. Cleveland (1993), "Discussion of Graphic Comparison of Several Linked Aspects: Alternative and Suggested Principles," Journal of Computational and Graphical Statistics, Vol. 2, No. 1, 41-48.

Carr, D.B. (1994a), "Using Gray in Plots", Statistical Computing and Statistical Graphics Newsletter, Vol 5, No. 2, pp 11-14.

Carr, D. B. (1994b), "Converting Tables into Plots." Technical Report No. 101, Center for Computational Statistics, George Mason University, Fairfax, VA.

Carr, D. B. (1994c), "Color Perception, the Importance of Gray and Residuals, on a Choropleth Map," Statistical Computing and Statistical Graphics Newsletter, Vol 5, No. 1, pp 16-20.

Frigge, M., Hoaglin, D. C. and Iglewicz, B. (1989), "Some Implementations of the Box Plot," The American Statistician, 43, 50-54.

Julesz, B., (1986), "Texton Gradients: The Texton Theory Revisited. Biological Cybernetics, No. 54, 245-251.

Kosslyn, S. M. (1994), Elements of Graphic Design, New York, NY: W. H. Freeman and Company.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978), "Variations of Box Plots," The American Statistician, 32, 12-16.

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

Tukey, J. W. (1989), "Data Based Graphics: Visual Display in the Decades to Come," Proceedings of the American Statistical Association Sesquicentennial Invited Paper Session, Alexandria, VA: The American Statistical Association, 366-381.

Tukey, J. W. (1993), "Graphic Comparisons of Several Linked Aspects: Alternative and Suggested Principles," Journal of Computational and Graphical Statistics, Vol. 2, No. 1, 1-33.

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

Tufte, E. R. (1990), Envisioning Information, Cheshire, CT: Graphics Press.

Daniel B. Carr
George Mason University
dcarr@galaxy.gmu.edu

NET SNOOPING

WWW viewers—Is there one for you?

by Mike Meyer

The Netscape Browser

In the last issue of the Newsletter I hinted at a software offering from Mosaic Communications. Well, the infobahn is racing ahead at full speed. The company changed its name to Netscape Communications (but kept the `mcom.com` Internet domain), released at least 3 beta versions of its software and won the praise or wrath of just about every denizen of the Internet. Even Time Magazine, available online at `http://www.timeinc.com`, in its December 5 1994 edition, has jumped into the fray, naming one of the Netscape developers, Marc Andreessen, to its list of 50 people with the "requisite ambition, vision and community spirit to help guide us in the new millennium". At least the electronic version of Time, from which I cut and pasted the above quote, mentions Andreessen. I haven't checked the printed version.

So what is all the noise about? Clearly not just a particular piece of software—but I will limit my comments to that.

Netscape is currently available at `ftp:ftp.mcom.com` and several mirror sites. There are versions for most popular Unix platforms (using X windows) and Macintosh and Windows versions. The user interface is very slick and remarkably similar across all three platforms, winning praise from many people. I regularly use the Mac and Unix versions and I have no difficulty switching between the two. The original Mosaic software, which by most measures would still be considered in its infancy, looks old and weary in comparison. Many of the people who praise the user interface also decry the memory requirements of the software. The Mac version requires 2Mb of memory and works best with 4Mb. Lots of people are spending lots of money on memory. Of course the people who complain about Netscape's memory requirements probably have not tried to install the latest commercial software from companies like Microsoft.

Netscape tries to do it all, and it largely succeeds.

Why should Netscape require so much memory? One reason is that Netscape tries to do it all, and it largely succeeds. Where other WWW viewers rely on external helper applications, especially to show non-inlined gifs