Variations of Box Plots

ROBERT McGILL, JOHN W. TUKEY, AND WAYNE A. LARSEN*

Box plots display batches of data. Five values from a set of data are conventionally used; the extremes, the upper and lower hinges (quartiles), and the median. Such plots are becoming a widely used tool in exploratory data analysis and in preparing visual summaries for statisticians and nonstatisticians alike. Three variants of the basic display, devised by the authors, are described. The first visually incorporates a measure of group size; the second incorporates an indication of rough significance of differences between medians; the third combines the features of the first two. These techniques are displayed by examples.

KEY WORDS: Box Plots; Exploratory data analysis; Graphical techniques.

1. Introduction

Box plots display batches of data (Tukey 1970, 1977). Five values from a set of data are conventionally used; the extremes, the upper and lower hinges[1] (quartiles), and the median. The basic configuration of the display is shown in Figure A. The technique has been used with considerable success in a diverse range of projects (cf. Cleveland, Dunn, and Terpenning 1976; Cleveland, Graedel, and Kleiner 1977; Cohen, Gnanadesikan, and Landwehr 1977; Kettenring et al. 1976). Inevitably, certain weaknesses came to light in particular cases; most frequently these were the result of inappropriate interpretation of the results rather than problems with the technique itself. In almost all cases, inclusion of additional available information in the display would have prevented the misinterpretation.

In an attempt to improve the basic display, three modified forms of box plots have been devised. The original version and these new variants are described by example in the following sections.

2. Basic Box Plot

Beginning in a positive vein, we first consider a case where the original method serves well and examine why this is the case. Figure B displays

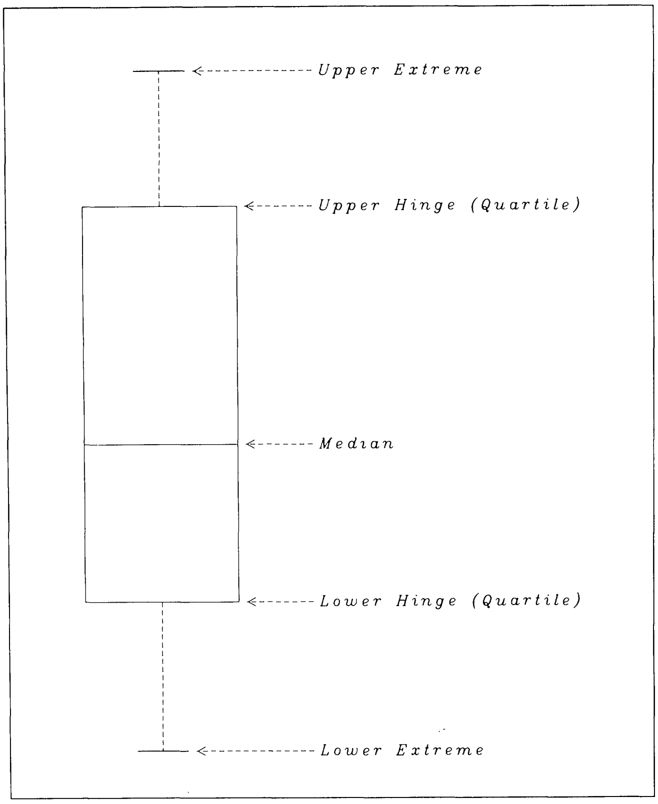

Figure A. Configuration of a Box Plot

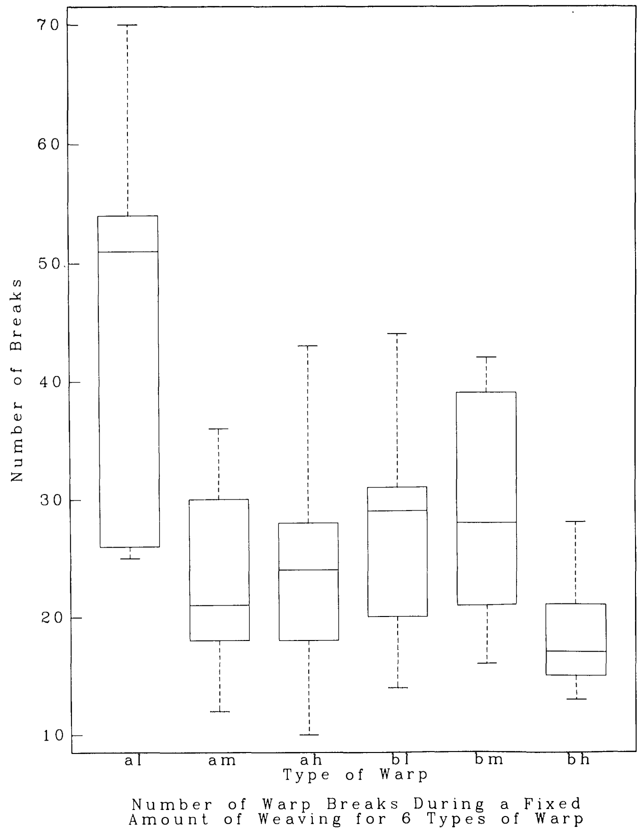

Figure B. Tippett's Warp Break Data

* Robert McGill is with Statistics and Data Analysis Research Department, Bell Laboratories, Murray Hill, NJ 07974. John W. Tukey is with Research-Communications Principles Division, Bell Laboratories, Murray Hill, NJ 07974, and Department of Statistics, Princeton University, Princeton, NJ 08540. Wayne A. Larsen is with Eyring Research Institute, Provo, UT 84601. This material was prepared in part in connection with research at Princeton University and United States Energy Research and Development Administration.

[1] The lower hinge is actually defined, for a sample of size n, as the $([(n + 1)/2] + 1)/2$th order statistic; [. . .] indicating the integer portion of the quotient. If the result is not an integer, the mean of the adjacent order statistics is used. The upper hinge is defined analogously.

Tippett's (1950) warp break data for six types of weaving warps. The characteristics of the data are easily seen, with type al appearing rather different from the other types.

Frequently, misinterpretation results because the viewer, particularly the nonstatistician, attempts to gain more information from the display than it contains. One might, for example, conclude that the overall median for all the types combined is about 26 or 27. While this conclusion is not justified based on the information contained, it is one that is often made. In this instance, it happens to also be correct—the actual overall median is 26.

In this example, two factors worked to the viewer's advantage. First, each group contained the same number of observations (specifically, nine). Second, the variance of each group, with the possible exception of the first, is moderately constant. In the absence of information to the contrary, the viewer will likely assume these facts (and, in this case, be correct). Next we examine a case where this is not so.

3. Variable-Width Box Plot

Figure C contains a regular box plot for another set of data. Here a single month's telephone bills for a group of Chicago residence customers is displayed. The data is subdivided by the number of years lived in the city. It should be emphasized that the plot correctly portrays the characteristics of the data displayed.[2] However, not everything known is shown.

On initial examination, perhaps the most striking feature is the pronounced drop seen in the last (over 15 year) group—the median is about \$13 while the medians of other groups are about \$20 or more. Returning to our rather naive user, he might conclude that the overall median for all groups combined is about \$21. This time the conclusion is not only unjustified, it is also grossly in error. The actual overall median is about \$14. What has gone wrong in this case?

The information available but not displayed is shown in the following tabulation. The number of customers in the various groups differ widely. In fact the ratio between the largest and the smallest is over 33:1.

Years	*Customers*
less than 1	11
1 to 2	17
3 to 5	26
6 to 10	35
11 to 15	29
over 15	368

[2] It will readily be admitted that plotting the data on a transformed scale (e.g., logs) might be preferable, although nontechnical viewers are often confused by this technique. Truncating the upper extremes on the display (not in computation) is another possible improvement. Both are illustrated later.

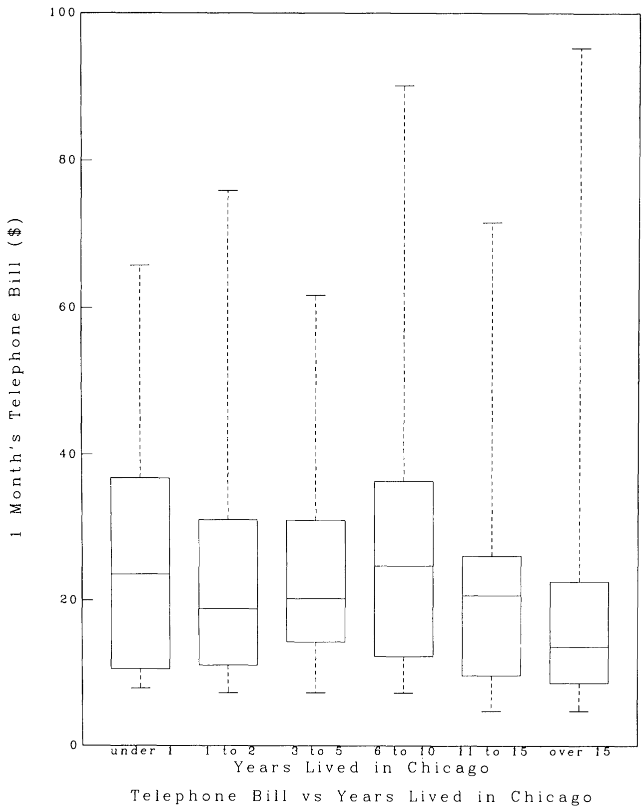

Figure C. Regular Box Plot

This large variation was obviously not deliberately introduced. Rather, it was the result of an unfortunate choice of group boundaries, made for the collection of the data, which caused over 75 percent of the customers to fall within one group.

Figure D shows a means of displaying this additional information—the variable-width box plot. Here the width of each box has been made proportional to the square root of the number of customers in the corresponding group.[3] The viewer's attention is immediately drawn to the size differences. (Obviously, a title clearly setting forth what has been done is definitely in order.) Since the additional information has been clearly displayed, a better appraisal of the data can be made and misinterpretations avoided.

4. Notched Box Plot

Returning to Figure C, the viewer might notice the surprisingly large (over \$5) difference in the medians of the first two boxes—classes which, intuitively at least, might be assumed to be rather similar. Were the variable-width plot examined, one might be led to doubt the significance of this difference. While it is evident that the number of customers in these groups is smaller than in other groups, actual size is not indicated. (Note that if the number of customers in

[3] The use of square root will be both intuitive and pleasing to the statistically inclined. However, as discussed in Section 6, other functions of group size may sometimes be more appropriate.

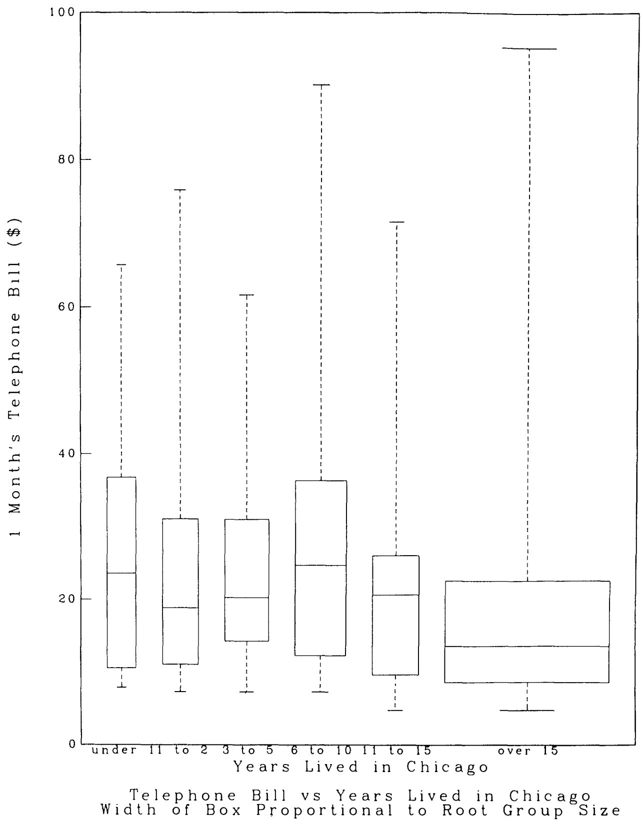

Figure D. Variable Width Box Plot

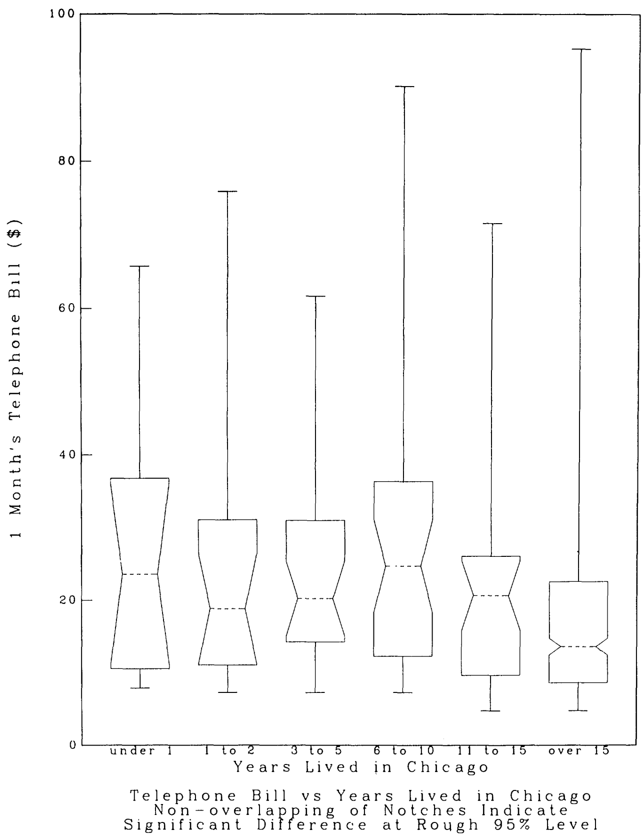

Figure E. Notched Box Plot

each group were 1,000 times larger, boxes of identical width would be produced.) Hence the viewer can determine confidence intervals around the medians only in a relative manner. Figure E, a notched box plot of the data, shows these explicitly. The notches surrounding the medians provide a measure of the rough significance of differences between the values. Specifically, if the notches about two medians do not overlap in this display, the medians are, roughly, significantly different at about a 95% confidence level.[4] Now the seemingly impressive difference previously noted is brought clearly into proper perspective—the difference is, in fact, not significant by our test. In fact, none of the differences seen in the first five boxes are significant.

It should be noted that the convention has been adopted that, should the notch lie outside either hinge, an unnotched box, plotted with dashed lines, is displayed for that group indicating low confidence in it. Experience has shown that few cases exist where this is not the appropriate strategy. These few, however, normally occur in cases where all notches protrude beyond a hinge, and/or one or more of the medians lies very near a hinge.

Figure F, using the warp data in Figure B, is such a case. All boxes would be dashed, and the median of type al is very near the upper hinge. Hence protruding notches have been plotted. Even so, little is gained except, perhaps, some slight confirmation of features already obvious.

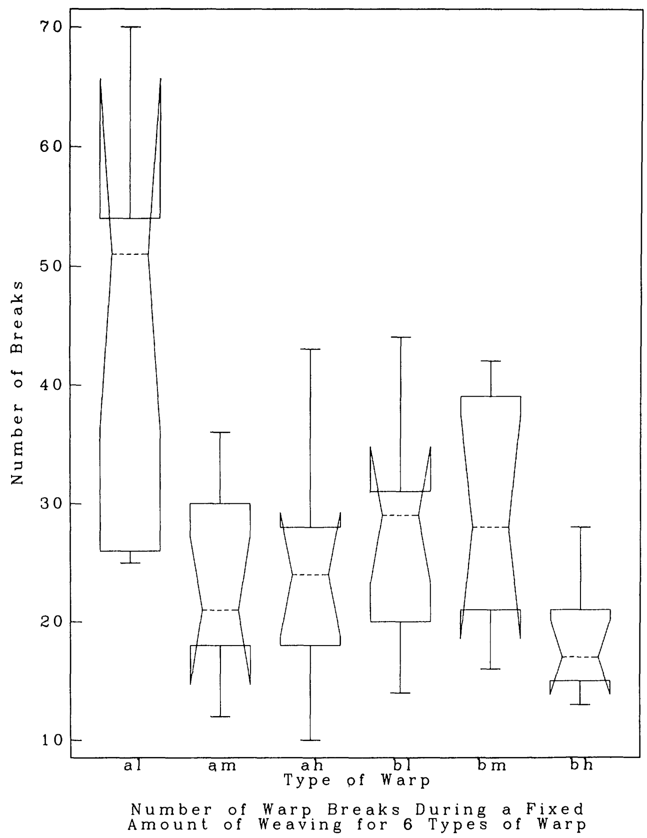

Figure F. Tippett's Warp Break Data—Notched

[4] Section 7 contains a description of the method used in determining the notch widths and suggests possible alternatives.

5. Variable-Width Notched Box Plot

In certain (perhaps many) cases, advantage will be gained by combining both the techniques described. Figures G through I contain such displays on the same telephone-bill data. In addition, the upper extremes have been truncated for plotting purposes only, and labeled appropriately in Figure G, and a log axis used in Figure H. Now both group size and confidence intervals can be seen simultaneously. As might be expected, the combination adds little in cases where a single technique (variable width or notches) would suffice. Here, due to the nature of the data, the combination does seem to provide an improvement.

Figure I contains the result of the next reasonable step in the analysis of this data. The first five groups have been combined, and boxes for under 15 years and over 15 years displayed. These two groups do appear significantly different—by at least about $4, the distance between the nearest edges of the notches.

Some might argue that the combination will normally contain considerable redundancy, since almost all reasonable measures of significant difference will be based in part on group size. While this fact is admitted, the question of whether such redundancy is necessarily undesirable may well be debated. In the final analysis, the user's personal preference is often the best criterion.

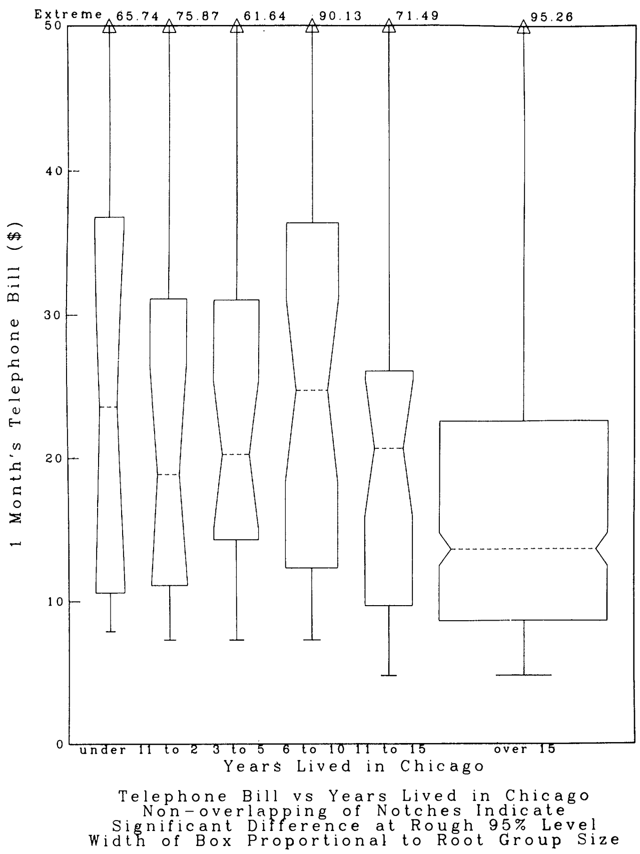

Figure G. Variable Width Notched Box Plot*
* Plot truncated at $50; extremes printed at top.

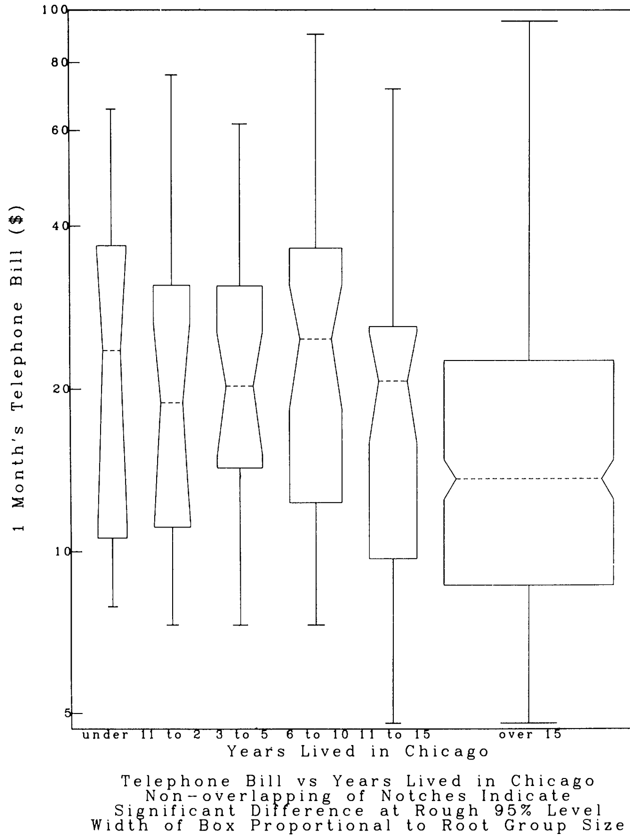

Figure H. Variable Width Notched Box Plot
NOTE: Y axis scale is logarithmic.

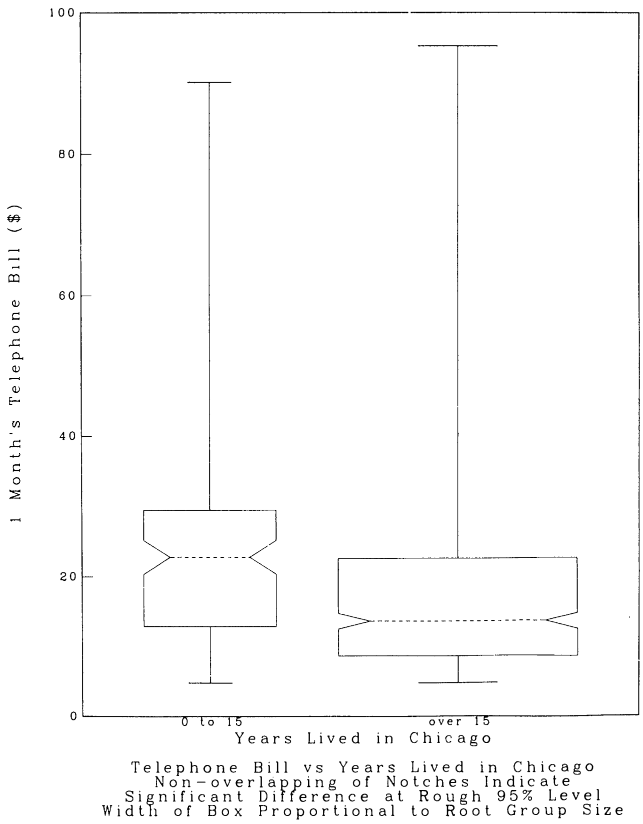

Figure I. Variable Width Notched Box Plot*
* The first five groups are combined.

6. Choice of Box Widths

In the examples shown in Figures D, G, H, and I, widths of the boxes have been made proportional to the square root of the group size. This choice was based on the fact that many variability measures, such as standard error, are proportional to the root of the group size. It is plausible to take this as the standard to be used unless other methods offer significant improvement.

One example would be when the actual intent is to display strata fractions. Since the use of a logit scale is sometimes preferred in these displays, box widths based on such a scale might be used here. Again, if the intent is to emphasize differences in group size, the use of square roots will clearly minimize the visual impact of differences, particularly in cases where the sizes are relatively similar. Here one might opt for using widths directly proportional to size. (However, limited experience in this area tends to indicate that this may overemphasize differences.) Again it must be stressed that whatever is done should be clearly indicated.

7. Choice of Notch Size

In notched box plots, one is, of course, faced with the question of how best to determine the widths of the notches. Many methods, both classical and nonparametric, might be considered. None will likely be best in all cases.

In Figures E, G, H, and I, the widths were computed from the midspread or interquartile range (R) of the data (a robust measure of spread), and the number of observations (N) for each group. The Gaussian-based asymptotic approximation (Kendall and Stuart 1967) of the standard deviation s of the median (M) is given by

$$s = 1.25R/1.35\sqrt{N} \qquad (7.1)$$

and can be shown to be reasonably broadly applicable to other distributions. An appreciation of why this is so can be obtained as follows. The asymptotic formula for the standard deviation of M is $1/2f_0\sqrt{N}$, where f_0 is the density at the population median. Also, R is a consistent estimate of the population interquartile range R_0, and $1/2R_0$ is the average density between the population quartiles. Thus for any distribution for which the middle portion is shaped approximately like a Gaussian, the ratio $f_0/(1/2R_0)$ will be close to the Gaussian value of 1.08. (This does not explain the rather remarkable result that for the (very skewed) one-sided exponential $2R_0f_0 = 1.10$.)

The notch around each median may then be calculated as

$$M \pm Cs, \qquad (7.2)$$

where C is a constant. Should one desire a notch indicating a 95 percent confidence interval about each median, $C = 1.96$ would be used. However, since a form of "gap gauge" which would indicate significant differences at the 95 percent level was desired, this was not done. It can be shown that $C = 1.96$ would only be appropriate if the standard deviations of the two groups were vastly different. If they were nearly equal, $C = 1.386$ would be the appropriate value, with 1.96 resulting in far too stringent a test (far beyond 99 percent). A value between these limits, $C = 1.7$, was empirically selected as preferable. Thus the notches used were computed as

$$M \pm 1.7(1.25R/1.35\sqrt{N}). \qquad (7.3)$$

Clearly, a variety of other choices, such as a single less conservative value (<1.7) or one dependent upon the data (chosen to compromise over the range of the ratios of the spreads involved), are possible and may be preferable in certain cases.

8. Display of Outliers

In many applications, and particularly for all but the least sophisticated viewers, special attention to outlying values may be desirable. This can be simply done by changing from box plots to schematic plots (Tukey 1977). Since the central box is the same on both, all the preceding discussion will apply.

9. Conclusion

Box plots have proven to be a most valuable tool in data analysis. The variants described—variable width boxes, notched boxes, and a combination of the two—provide additional information on the display. Hopefully, these not only facilitate interpretation and provide additional insight into the data but also lessen the possibility of misinterpretations due to unwarranted assumptions.

[*Received January 26, 1977. Revised September 13, 1977.*]

References

Cleveland, W. S., Dunn, D. M., and Terpenning, I. (1976), "SABL—Resistant Seasonal Adjustment Procedure with Graphical Methods for Interpretation and Diagnosis," *Proceedings of the NBER-CENSUS Conference on Seasonal Analysis of Economic Time Series,* U.S. Bureau of the Census: Washington, D.C.

———, Graedel, T. E., and Kleiner, B. (1977), "Urban Formaldehyde: Observed Correlation with Source Emissions and Photochemistry," *Atmospheric Environment,* 11, pp. 357–360.

Cohen, A., Gnanadesikan, R., Kettenring, J. R., and Landwehr, J. M. (1977), "Methodological Developments in Some Applications of Clustering," *Proceedings of the Symposium on Applications of Statistics,* ed. P. R. Krishnaiah, North-Holland Publishing Co.

Kendall, M. G., and Stuart, A. (1967), *The Advanced Theory of Statistics, Vol. 1,* 2nd ed., Ch. 14, New York: Hafner Publishing Co.

Kettenring, J. R., Rogers, W. H., Smith, M. E., and Warner, J. L. (1976), "Cluster Analysis Applied to the Validation of Course Objectives," *Journal of Educational Statistics,* 1, No. 1, 39–57.

Tippett, L. H. C. (1950), *Technological Applications of Statistics,* New York: John Wiley and Sons, p. 106.

Tukey, J. W. (1970), *Exploratory Data Analysis (Limited Preliminary Edition), Vol. 1,* Ch. 5, Reading, Mass: Addison-Wesley Publishing Co.

——— (1977), *Exploratory Data Analysis* (First Edition), Reading, Mass: Addison-Wesley Publishing Co.