The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Some Implementations of the Boxplot

Michael Frigge , David C. Hoaglin & Boris Iglewicz

To cite this article: Michael Frigge , David C. Hoaglin & Boris Iglewicz (1989) Some Implementations of the Boxplot, The American Statistician, 43:1, 50-54

To link to this article: https://doi.org/10.1080/00031305.1989.10475612

STATISTICAL COMPUTING

This department includes the two sections New Developments in Statistical Computing and Statistical Computing Software Reviews; suitable contents for each of these sections are described under the respective section heading. Articles submitted for the department, outside the two sections, should not be highly technical and should be relevant to the teaching or practice of statistical computing.

Some Implementations of the Boxplot

MICHAEL FRIGGE, DAVID C. HOAGLIN, and BORIS IGLEWICZ*

An increasing number of statistical software packages offer exploratory data displays and summaries. For one of these, the graphical technique known as the boxplot, a selective survey of popular software packages revealed several definitions. These alternative constructions arise from different choices in computing quartiles and the fences that determine whether an observation is "outside" and thus plotted individually. We examine these alternatives and their consequences, discuss related background for boxplots (such as the probability that a sample contains one or more outside observations and the average proportion of outside observations in a sample), and offer recommendations that lead to a single standard form of the boxplot.

KEY WORDS: Exploratory data analysis; Outliers; Quartiles; Statistical software.

1. INTRODUCTION

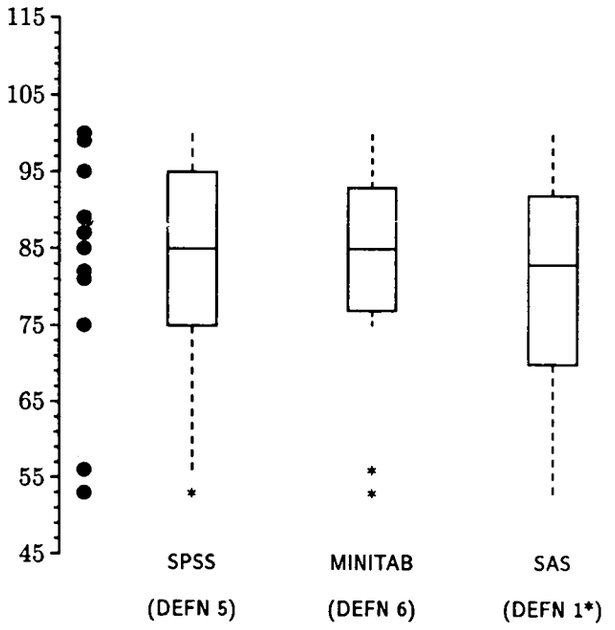
The boxplot, a popular univariate data display developed by John W. Tukey (1970, 1977), is available in many statistical software packages. Velleman and Hoaglin (1981) discussed this display and its construction in detail. We have studied how the boxplot, as an example of the many techniques from exploratory data analysis and other areas of statistics, is actually implemented in several heavily used packages. In addition, we discuss some key properties of the boxplot in more detail. We show that some major statistical packages use nonstandard definitions of the hinges and alternative values of the constant that controls the fences. Thus the resulting boxplots can vary in shape and in the number of observations identified as "outside." We provide suggestions for reducing this diversity.

We have not attempted to make an exhaustive survey of available software, especially for microcomputers. Instead, we restrict our comments to Minitab, S, SAS, SPSS, Statgraphics, and Systat, which were conveniently available. [See Ryan, Joiner, and Ryan (1985) for Minitab; Becker and Chambers (1984) for S; SAS Institute Inc. (1985); SPSS, Inc. (1986); STSC, Inc. (1985) for Statgraphics; and Wilkinson (1986) for Systat.] All results are based on our hands-on investigation. We used the mainframe versions of the first four packages and the micro versions of the last two.

As a first step toward understanding the boxplots that statistical packages acutally produce, we consider an example. Throughout, we are concerned with the numerical elements of the boxplot, rather than with graphical style. For example, the definition of the display does not prescribe the dimension of the box perpendicular to the data axis. We ordinarily prefer to avoid boxes that are very thin or very fat; but we leave this choice to the implementer and, when feasible, to the user. Similarly, an implementation may draw the box or assemble it from standard symbols in the ASCII character set.

For the example we constructed a batch consisting of the following 11 observations: 53, 56, 75, 81, 82, 85, 87, 89, 95, 99, 100. Figure 1 shows a dot plot of the batch and the boxplots that three packages produce for that batch under certain options. If we approach these as boxplots for three separate batches, we might say that they have nearly the same median, the middle one has somewhat less spread and two low outside observations, the one on the left has one low outside value, and the one on the right has no outside values. These three boxplots come from SPSS, Minitab, and one option in SAS, respectively. Notice that, for this particular option (not the default) in SAS, even the median can differ from the conventional definition, which takes the middle observation when n is odd and averages the middle two observations when n is even. This example illustrates that displays delivered as boxplots can give rather different impressions of the same batch of data. As long as the displays use the same definition of outside, such differences must become small as the sample size increases. Because the boxplot does not show n, however, users should keep the possible differences in mind when displaying small samples. The next three sections discuss the reasons for these variants.

*Michael Frigge is a graduate student, Department of Statistics, University of Chicago, Chicago, IL 60637. He completed a master's degree at Temple University while working on this article. David C. Hoaglin is Research Associate, Department of Statistics, Harvard University, Cambridge, MA 02138. Boris Iglewicz is Professor, Department of Statistics, Temple University, Philadelphia, PA 19122. This work was supported in part by U.S. Army Research Office Contract DAAG29-85-K-0262 with Harvard University. The authors are grateful to two referees for helpful comments. An earlier version of this article appeared in *Computer Science and Statistics: Proceedings of the 19th Symposium on the Interface.*



* SAS uses definition 4 as the default.

Figure 1. Boxplots for One Sample (shown at left) Produced by Three Statistical Packages.

2. BOXPLOT CONSTRUCTION

A boxplot aims to summarize a batch of data by displaying several main features, as illustrated by the hypothetical boxplot in Figure 2. It derives its name from the rectangular box, which locates the middle half of the batch. A line across the box shows the position of a typical central value. These two features, ordinarily shown by solid lines, attract the majority of a viewer's attention—as they should, because summaries of univariate data focus first on location and spread.

From the two ends of the box, dashed lines extend outward to the two adjacent values, the outermost observations that are not extreme enough to be flagged as outside by an exploratory rule of thumb. The display shows any outside observations individually so that they will attract attention and receive further examination as potential outliers.

In preparation for discussing alternatives, we now review the construction details of these features. The center of the boxplot is usually the sample median, although other choices have been suggested for special applications. For example, Iglewicz and Hoaglin (1987) used the F-mean (the average of the points inside and on the boundary of the box) in a quality control chart. Some implementations introduce a further single character to show the position of the sample mean. Surprisingly, one may encounter definitions of the median that differ from the conventional one. One example takes the $(n/2)$th ordered observation when n is even and averages the $(n-1)/2$th and $(n+1)/2$th observations when n is odd. Still, most software packages use the standard definition.

The location of the sample quartiles or the fourths involves considerably greater variety. As introduced by Tukey (1977, p. 39), the boxplot has the ends of the box at the hinges or fourths, whose definition we review in the next section. Some implementations, however, use the sample quartiles, calculated according to one of several definitions. For example, the UNIVARIATE procedures in SAS allows users to choose among five definitions for sample percentiles and then uses the resulting quartiles for the boxplot if a user chooses the PLOT option. As it happens, none of the five definitions of quartiles coincides with the fourths. This situation seems unfortunate, because the fourths (or quartiles) play a key role in the construction of the boxplot.

Thus different definitions can occasionally lead to substantially different impressions of the same batch of data. Someone who has studied the standard boxplot [e.g., in Koopmans (1981) or Velleman and Hoaglin (1981)], and who assumes that the various software packages follow the standard definition, may run some risk of being misled.

The quartiles take on extra importance because they are the basis for the fences in the rule for flagging potential outliers. If Q_1 and Q_3 are the lower quartile and the upper quartile, respectively, then the fences lie at $Q_1 - k(Q_3 - Q_1)$ and $Q_3 + k(Q_3 - Q_1)$, customarily with $k = 1.5$. Thus the choice of definition for the quartiles can affect the number of observations shown as outside. In addition, a change in the value of k can have an even greater impact, as we discuss in Section 4.

A special version of the boxplot, introduced by Cleveland (1985, pp. 129–131), substitutes the 10th and 90th percentiles of the sample in place of the fences. This innovation should be helpful when one uses a boxplot to summarize an entire population, but (as we explain in Sec. 4) it will show too many observations individually when displaying a sample.

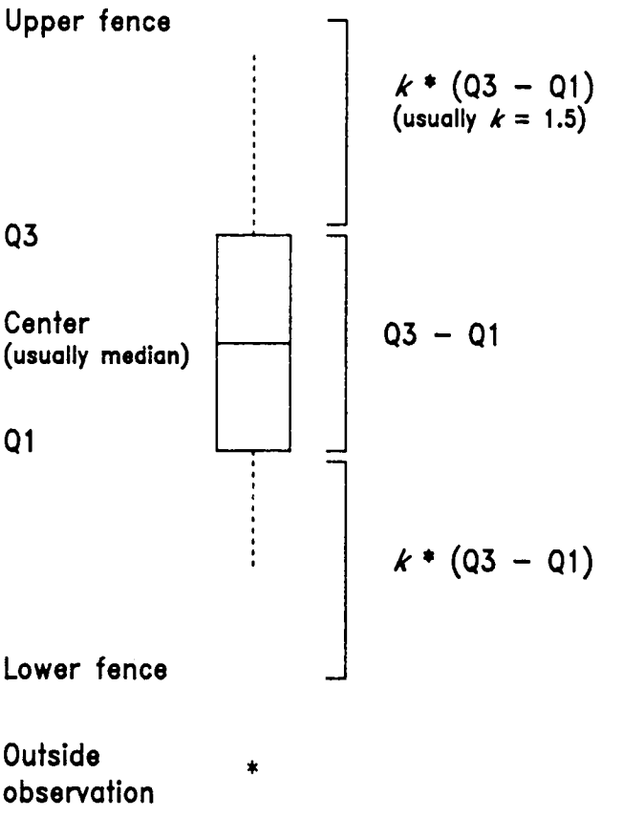


Figure 2. A Typical Boxplot.

3. QUARTILES

Several definitions of quartiles are in use. As one indication of the diversity, Freund and Perles (1987) discussed three versions of interpolation and introduced a scheme for dividing a sample into four parts. We list eight choices in Figure 3; the first five come from SAS, Definition 6 (used by Minitab and Systat) comes from Tukey (1977), Definition 7 is recommended by Hoaglin and Iglewicz (1987), and Definition 8 (in the context of the boxplot) appears in Cleveland (1985). Definition 4 (also used by S) is the default definition in the UNIVARIATE procedure of SAS, and Definition 5 is the one used by SPSS and Statgraphics. In Figure 3 we also define the lower quartile in terms of the ordered observations $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.

The literature of exploratory data analysis introduced the term *fourth* (earlier, *hinge*) as a reminder that Definition 6 has an aim slightly different from previous definitions of quartile. In fact, strictly speaking, Definitions 6 and 7 yield fourths. The difference lies in the way that one assigns tail areas to order statistics. If $X_{(i)}$ denotes the ith order statistic in a random sample of n, then $(i - \frac{1}{3})/(n + \frac{1}{3})$ is a good approximation to the tail area corresponding to the median of the sampling distribution of $X_{(i)}$. That is, as discussed by Hoaglin (1983),

$$\text{Med}(X_{(i)}) \approx F^{-1}\left[\frac{i - \frac{1}{3}}{n + \frac{1}{3}}\right],$$

where F is the underlying cdf. Definition 7 solves $((j+g) - \frac{1}{3})/(n + \frac{1}{3}) = \frac{1}{4}$ for $j + g$ (where j is an integer and $0 \le g < 1$) and obtains the lower fourth as $(1-g)x_{(j)} + gx_{(j+1)}$. Definition 6 merely simplifies the result for hand calculation (its $j + g$ exceeds that from Def. 7 by either $\frac{1}{3}$ or $\frac{1}{12}$, depending on n). For example, if $n = 11$, Definition 7 gives $j + g = 3 + \frac{1}{6}$, whereas Definition 6 gives $j + g = 3 + \frac{1}{2}$.

Definitions 1–5 and 8 are the result of substituting $p = \frac{1}{4}$ into the corresponding definitions of the $100p$th percentile. In this framework one obtains the upper quartiles by similarly substituting $p = \frac{3}{4}$. Definitions 6 and 7, however, give only the depth of the fourth from the nearer end of the ordered batch. If d is that depth, one is supposed to take $x_{(d)}$ as the lower fourth and $x_{(n+1-d)}$ as the upper fourth, interpolating as necessary. This approach explicitly treats the two ends of the batch symmetrically. Among the definitions based on percentiles, only Definitions 4, 5, and 8 satisfy the symmetry condition $\text{rank}(Q_3) = n+1 - \text{rank}(Q_1)$. This fact comes as no surprise, because the other basic definitions are intended to serve other purposes, but it implies that Definitions 1, 2, and 3 should not be used to obtain quartiles for a boxplot.

Among Definitions 4–8, no two give identically the same depth. The depths based on Definitions 4, 7, and 8 increase linearly with sample size, whereas Definitions 5 and 6 yield depths that stay constant or increase in jumps of .5 as n increases. Figure 4 plots the depth of the quartiles from $n = 4$ to $n = 20$ for the commonly used Definitions 4, 5, and 6 plus Definition 7. The depths based on Definition 7 follow a smooth compromise line, one aim of this definition. However, to avoid giving positive weight to the end observations when $n = 5$ or $n = 6$ (which would make it nearly impossible for them to be flagged as outside), we prefer to modify Definition 7 by taking the fourths at depth 2 in batches of 5 or 6.

In summary, among the numerous choices for the depth of the quartiles or fourths in a boxplot, Minitab and Systat use Definition 6, SPSS and Statgraphics use Definition 5, and S and the default in SAS use Definition 4. We feel that

Definition 1: Weighted Average at $x_{(n/4)}$
$Q_1 = (1-g)x_{(j)} + gx_{(j+1)}$.

Definition 2: Observation Numbered Closest to $n/4$
$Q_1 = x_{(i)}$, where i is the integer part of $n/4 + \frac{1}{2}$.

Definition 3: Empirical Distribution Function
$Q_1 = x_{(j)}$ if $g = 0$; $Q_1 = x_{(j+1)}$ if $g > 0$.

Definition 4: Weighted Average Aimed at $x_{((n+1)/4)}$ (used in S, default in SAS UNIVARIATE)
$Q_1 = (1-g)x_{(j)} + gx_{(j+1)}$, where $(n+1)/4 = j + g$.

Definition 5: Empirical Distribution Function With Averaging (used in SPSS and Statgraphics)
$Q_1 = (x_{(j)} + x_{(j+1)})/2$ if $g = 0$; $Q_1 = x_{(j+1)}$ if $g > 0$.

Definition 6: Standard Fourths or Hinges (used in Minitab and Systat)
$Q_1 = (1-g)x_{(j)} + gx_{(j+1)}$, where $[(n+3)/2]/2 = j + g$ (note that $g = 0$ or $g = \frac{1}{2}$).

Definition 7: Ideal or Machine Fourths
$Q_1 = (1-g)x_{(j)} + gx_{(j+1)}$, where $n/4 + 5/12 = j + g$.

Definition 8: Weighted Average Aimed at $x_{(n/4+.5)}$
$Q_1 = (1-g)x_{(j)} + gx_{(j+1)}$, where $n/4 + \frac{1}{2} = j + g$.

Figure 3. Eight Definitions of the Lower Quartile Q_1 in Terms of the Ordered Observations $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. Unless otherwise stated, $n/4 = j + g$, j is an integer, and $0 \le g < 1$; $[x]$ denotes the largest integer that does not exceed x. Definitions 1–5 are from SAS Institute Inc. (1985, pp. 1186–1187), Definition 6 is from Tukey (1977, p. 33), Definition 7 is from Hoaglin and Iglewicz (1987), and Definition 8 is from Cleveland (1985, p. 130).

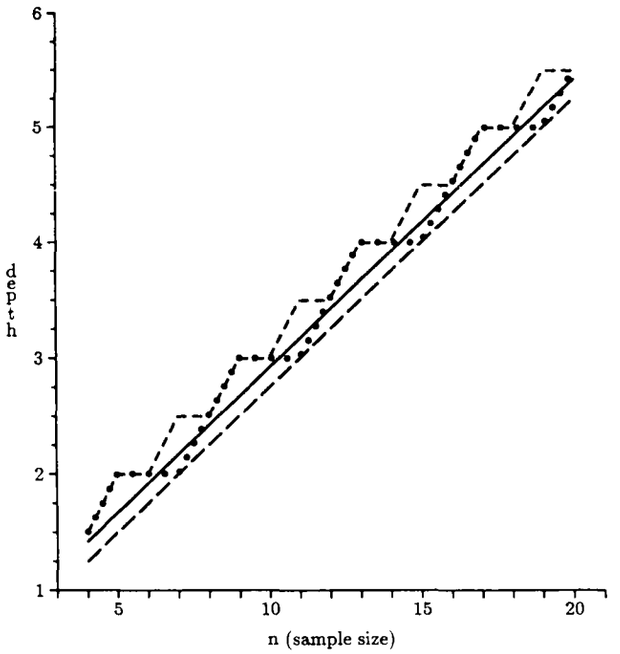


Figure 4. Depth of Lower Quartile for Selected Sample Sizes, Using Definitions 4–7: ———, Definition 4; ···, Definition 5; ---, Definition 6; ——, Definition 7.

this lack of standardization leads to considerable confusion. Although Definition 7 may eventually become standard, for now we recommend the current standard, Definition 6. In SAS we would use Definition 4 (when $n \geq 7$). At present the low resolution of printer-plot implementations may conceal some of the differences among definitions, but advances in hardware and software promise to make high-resolution displays more widely available.

4. VALUES OF k FOR FENCES

Several values of k have been recommended for use in the rule of thumb for flagging observations as outside. Tukey (1977, p. 44) defined the inner fences by $k = 1.5$ and the outer fences by $k = 3.0$. An earlier pair of thresholds, preserved by McNeil (1977), uses $k = 1.0$ and $k = 1.5$. Yet another version (Tukey 1972) has $k = 1.0$ and $k = 2.0$. The chronological order of these versions agrees with our understanding that they developed in response to accumulating experience: First $k = 1.0$ and $k = 1.5$, then $k = 1.0$ and $k = 2.0$, and finally $k = 1.5$ and $k = 3.0$. SPSS bases its boxplot on McNeil's text. Ingelfinger, Mosteller, Thibodeau, and Ware (1983), on the other hand, used $k = 2.0$.

Which of these values should be used, and for which purpose? Hoaglin, Iglewicz, and Tukey (1986) provided some guidance by studying the some-outside rate per sample. This measure equals the probability that a random sample of n contains one or more outside observations. For Gaussian data they denoted this rate by $1 - B(k, n)$. Table 1 gives the value of $1 - B(k, n)$ at selected values of k and n, when the fourths are obtained according to the standard definition (Def. 6). As Hoaglin et al. (1986) explained, this definition leads to values of $1 - B(k, n)$ whose relation to n (for fixed k) depends somewhat on the remainder n mod 4. Here we focus on $n = 10, 20, 30, 50$, and 100 and consider the behavior of $1 - B(k, n)$ as n becomes large. For any fixed k, $1 - B(k, n)$ must approach 1 as n becomes large, because a large-enough sample is virtually certain to contain at least one observation outside the fences. The column for $k = 1.0$ in Table 1 clearly exhibits this behavior, and the column for $k = 1.5$ shows less dramatic, but still definite, evidence of it.

Because the rule with $k = 1.0$ often flags at least one observation in more than 50% of samples of uncontaminated Gaussian data, we regard it as unsatisfactory. For $5 \leq n \leq 20$ the basic rule ($k = 1.5$) has a some-outside rate per sample of about 25%. This seems acceptable when we are proceeding in an exploratory mode, because then we are prepared to give further attention to a moderate number of observations. Indeed, as Hampel (1985) pointed out, in this mode it is more important not to miss any potential outlier than to avoid casting doubt on good observations. The rule with $k = 2.0$ comes close to the 10% rate that one might use in outlier detection (and rejection). The table entries for $k = 3.0$ show why such outer fences are suitable for identifying highly unusual data: Only a small percentage of Gaussian samples contains one or more observations beyond the outer fences.

In addition to the some-outside rate per sample, Hoaglin et al. (1986) also defined the outside rate per observation as the average fraction of outside observations. This rate has sometimes been used to guide or document choices of k. For example, in describing the SPLOT procedure the SAS *Supplemental User's Guide* states that "values less than ($Q1 - 1.5*Q$RANGE) or greater than ($Q3 + 1.5*Q$RANGE) . . . occur about one in two hundred for normal samples" (SAS Institute Inc. 1983, p. 341). This statement paraphrases a remark of Tukey (1970, p. 5-21). Since we now have considerable information (for $k = 1.5$ and the standard definition of the fourths) for small to moderate Gaussian samples, as well as the corresponding population value, we can supply a more accurate description. In fact, this rate varies substantially with n. From the more detailed discussion of Hoaglin et al. (1986) we obtain these illustrative approximate values: $n = 10$, rate 1 in 35; $n = 20$, rate 1 in 60; $n = 50$, rate 1 in 87; $n = \infty$, rate 1 in 143. Thus the population value is a poor indicator of the behavior in small samples. For Cleveland's version, by comparison, the outside rate per observation is 1 in 5 at all sample sizes. We customarily put greater emphasis on the some-outside rate per sample (an experimentwise error rate) than on the outside rate per observation.

Table 1. The Some-Outside Rate per Sample in Gaussian Samples, $1 - B(k, n)$, for Selected Values of k and n

	k			
n	1.0	1.5	2.0	3.0
10	.424	.198	.094	.026
20	.577	.232	.082	.011
30	.705	.284	.094	.008
50	.837	.365	.094	.004
100	.967	.523	.115	.003

NOTE: These entries come from simulation studies of the exploratory rule based on the standard definition of the fourths (Def. 6).

5. SUMMARY AND RECOMMENDATIONS

By taking the boxplot as an illustrative example, we have seen that implementations of a procedure can vary substantially from package to package. This means that users must take special care to understand which algorithms have been used. In particular, differences among boxplots in Minitab, Systat, S, SAS, SPSS, and Statgraphics arise from differences in the definition of quartiles.

Besides the quartiles, the fences also need to be standardized. Substantial numerical evidence now indicates that $k = 1.0$ is too small. We prefer the standard $k = 1.5$, which seems satisfactory for exploratory purposes. As options, we can see reasons for offering $k = 2.0$, as well as the outer fences at $k = 3.0$.

We believe that the benefit to users of consistency across packages outweighs the minor additional calculation that some packages might have to make in order to deliver a standard boxplot. Velleman and Hoaglin (1981) provided programs in FORTRAN and BASIC. If such a boxplot cannot easily be installed as the default, at least packages could offer it is an option. Then users can concentrate on the data and their analysis, without having to allow for variations in the definitions.

We also urge that documentation state more carefully what is being computed and explain the reasons for nonstandard features. For a technique as straightforward as the boxplot, users should readily be able to interpret the output and verify the formulas.

We recognize that new statistical techniques may undergo further development and refinement, even after evidence of their merit has become strong enough to justify their inclusion in major software systems. Rather than add to existing lags in statistical technology, we favor provisional implementations of new techniques at a relatively early stage, with the understanding that some aspects of the implementation may change as related research and user feedback accumulate.

[*Received August 1987. Revised July 1988.*]

REFERENCES

Becker, R. A., and Chambers, J. M. (1984), *S: An Interactive Environment for Data Analysis and Graphics*, Belmont, CA: Wadsworth.

Cleveland, W. S. (1985), *The Elements of Graphing Data*, Monterey, CA: Wadsworth.

Freund, J. E., and Perles, B. M. (1987), "A New Look at Quartiles of Ungrouped Data," *The American Statistician*, 41, 200–203.

Hampel, F. R. (1985), "The Breakdown Points of the Mean Combined With Some Rejection Rules," *Technometrics*, 27, 95–107.

Hoaglin, D. C. (1983), "Letter Values: A Set of Selected Order Statistics," in *Understanding Robust and Exploratory Data Analysis*, eds. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: John Wiley, pp. 33–57.

Hoaglin, D. C., and Iglewicz, B. (1987), "Fine-Tuning Some Resistant Rules for Outlier Labeling," *Journal of the American Statistical Association*, 82, 1147–1149.

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.

Iglewicz, B., and Hoaglin, D. C. (1987), "Use of Boxplots for Process Evaluation," *Journal of Quality Technology*, 19, 180–190.

Ingelfinger, J. A., Mosteller, F., Thibodeau, L. A., and Ware, J. H. (1983), *Biostatistics in Clinical Medicine*, New York: Macmillan.

Koopmans, L. H. (1981), *An Introduction to Contemporary Statistics*, Boston: Duxbury Press.

McNeil, D. R. (1977), *Interactive Data Analysis*, New York: John Wiley.

Ryan, B. F., Joiner, B. L., and Ryan, T. A. (1985), *Minitab Handbook* (2nd ed.), Boston: Duxbury Press.

SAS Institute Inc. (1983), *SUGI: Supplemental User's Guide*, Cary, NC: Author.

——— (1985), *SAS User's Guide: Basics* (Version 5 ed.), Cary, NC: Author.

SPSS, Inc. (1986), *SPSS-X, User's Guide* (2nd ed.), New York: McGraw-Hill.

STSC, Inc. (1985), *Statgraphics: Statistical Graphics System by Statistical Graphics Corporation*, Rockville, MD: Author.

Tukey, J. W. (1970), *Exploratory Data Analysis* (limited preliminary ed.), Reading, MA: Addison-Wesley.

——— (1972), "Some Graphic and Semigraphic Displays," in *Statistical Papers in Honor of George W. Snedecor*, ed. T. A. Bancroft, Ames: Iowa State University Press, pp. 293–316.

——— (1977), *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Velleman, P. F., and Hoaglin, D. C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis*, Boston: Duxbury Press.

Wilkinson, L. (1986), *Systat: The System for Statistics*, Evanston, IL: Systat, Inc.