In Reply Refer To:
WGS-Mail Stop 410 June 19, 1989

BRANCH OF SYSTEMS ANALYSIS TECHNICAL MEMORANDUM NO. 89.01

Subject: STATISTICS -- New Series of Technical Briefing Papers

Attached is the first in a new series of briefing papers on statistical topics of interest in hydrology. The topics chosen will span the disciplines of water quality, surface water and ground water; discipline-specific topics are already addressed in technical memos from the separate Offices. Each paper will present topics which are applicable to ongoing work within the Division, and will represent activities either of the Branch of Systems Analysis or related work in the field of applied statistics. They attempt only to survey these topics, and the appropriate references should be sought for more detail. Professionals within the Division should become familiar with the contents of these memos.

This first paper describes the use and construction of boxplots. Boxplots more clearly display the characteristics of a data set than do either histograms or 'dot and line' plots of means with error bars. They are particularly useful for comparing the characteristics of several data sets. Boxplots are increasingly being used in WRD reports, and the appendix of this paper should clarify the methods by which various software packages construct boxplots. Later papers will discuss recent advances in techniques of load estimation, treatment of data below the detection limit, and trend analysis.

Dennis Helsel
Acting Chief,
Branch of Systems Analysis

Attachment

This memorandum does not supersede any previous WRD technical memorandum.

Key words: Statistics, Data Analysis, Boxplots

Distribution: w / attachment: S, PO, FO
w / o attachment: A, B

# Boxplots

A Graphical Method for Data Analysis

by Dennis Helsel

This briefing paper describes a graphical method which greatly enhances a scientist's understanding of data, and for illustrating those data in reports -- boxplots. The objectives of this paper are to introduce this procedure, clarify terminology and definitions, provide locations of existing USGS computer code, and list references for more detail.

## Introduction and References

Boxplots graphically summarize and portray the characteristics of one or more data sets. They are alternatives to histograms for this purpose, and are particularly useful when comparing multiple data sets. They display

1) the center of the data (the median--the center line of the box)
2) the variation or spread (interquartile range--the box length)
3) the skewness (size of box halves, length of whiskers)
4) presence or absence of unusual values ("outside" values).

A good reference for graphical procedures, including boxplots, is Chambers and others (1983). Velleman and Hoaglin (1981) also provide good descriptions of boxplots and other exploratory data analysis procedures. The original boxplot reference, which is difficult to read, is Tukey (1977). Variations on the original plot, of which there are several, are given by McGill and others (1978). Three commonly used versions of the boxplot, all of which have already appeared in WRD reports, are described below. Any of the three may appropriately be called a boxplot.

Simple boxplot

The simple boxplot was originally called a "box-and-whisker" plot by Tukey (1977). It consists of a center line (the median) splitting a rectangle defined by the upper and lower quartiles, or fourths (see appendix). Whiskers are lines drawn from the ends of the box to the maximum and minimum of the data (Figure 1a). Simple boxplots have been used in WRD reports by Hren and others (1984), and Chen and Druliner (1987), among others.

Standard Boxplot

Tukey's "schematic plot" has become the most commonly-used version of a boxplot (figure 1b). With this standard boxplot, unusual values are distinguished from the rest of the plot. The box is as defined above. However, the whiskers are shortened to extend only to the last observation within one step beyond either end of the box. A step equals 1.5 times the length of the box (1.5 times the interquartile range). Observations farther than one step from the box in either direction are plotted individually, usually with an asterisk. Observations farther than two steps from the box are additionally distinguished by using a different symbol, often a small circle. Standard boxplots have

been used in WRD reports by Schertz and Hirsch (1985) and Grady and Weaver (1987), among others.

Truncated Boxplot

In a third version of the boxplot (figure 1c), the whiskers are drawn only to the 90th and 10th percentiles of the data set. The largest 10 percent and smallest 10 percent of the data are not shown. The National Water Summary has adopted this version for its portrayal of data (Moody and others, 1988). This version could easily be confused with the simple boxplot, as no data appear beyond the whiskers, and should be clearly defined as having eliminated the most extreme 20 percent of data. It should be used only when the extreme 20 percent of data are not of interest.

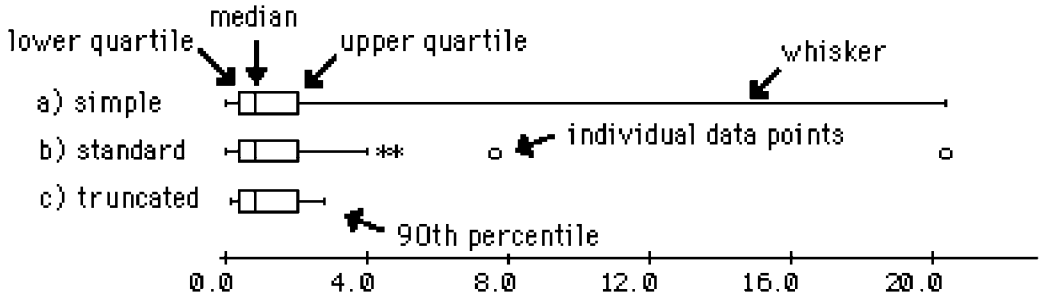



Figure 1. Three versions of the boxplot

**Uses of Boxplots**

The primary use of boxplots is for comparison between data sets. As an example, Figure 2 shows boxplots for dissolved solids concentrations going downstream along the Colorado River. The general trend of increasing concentrations is illustrated by the generally increasing location of the center line of the box, the median concentration. The box lengths add the additional information that variability in concentration (the interquartile range) also changes going downstream. In particular, a large drop in both concentration and variability occurs downstream of Glen Canyon Dam. The boxes also show that most of the data are approximately symmetric, as the top and bottom halves of the boxes are about the same lengths, as are the lengths of the top and bottom whiskers for each box. One unusually low value occurs at Aqueduct; rather than deleting it, perhaps it represents some unusual event. That single point would have distorted a 'dot and line' plot of mean and standard deviation, lowering the mean and greatly inflating the standard deviation. Thus this one point would have incorrectly implied that the entire data set for that site had characteristics different than they actually were.

A second use of boxplots is for a visual determination of whether data fit the assumptions of a statistical test procedure. The change in variability, for example, will cause difficulties for a linear regression of concentration versus distance, as linear regression assumes constant variance. Similarly, a t-test for a change in concentration between Cisco and Lees Ferry stations (above versus below Glen Canyon Dam) would be inappropriate without adjustment for differences in variance, as the t-test also assumes constant variance. Finally, the data at all but one site (Aqueduct) appear approximately symmetric with only a few outliers. Therefore the assumption of normality for each data group (station) that is required by a t-test or analysis of variance is probably not strongly violated.

When evaluating software, the ability to produce side-by-side boxplots is important. Although most commercial statistical software will produce these, SAS presently does not.

## Availability of Computer Code

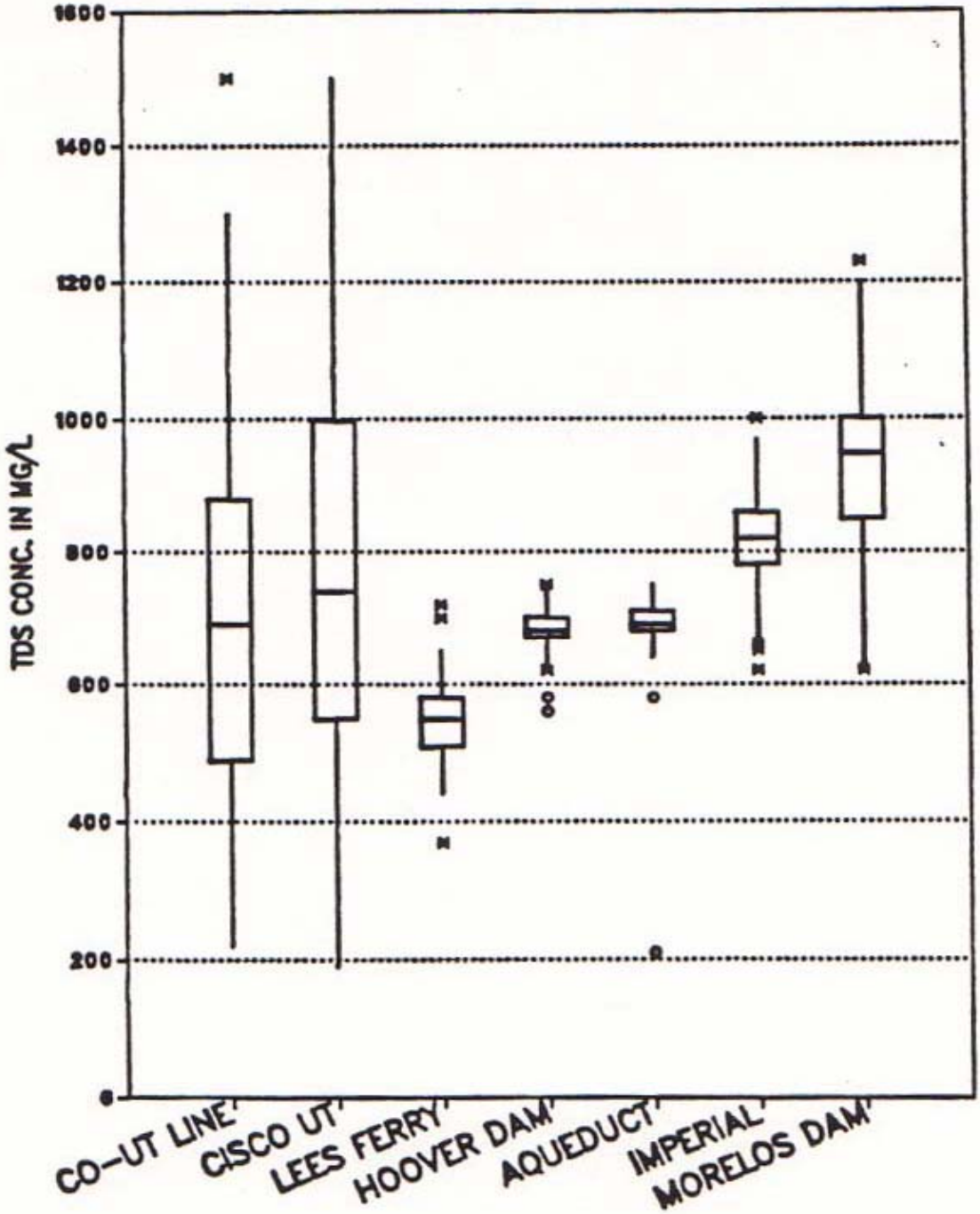
A Fortran program to produce side-by-side boxplots written several years ago within the Branch of Systems Analysis is available on the RVARES Prime computer. It resides in a directory accessible to the FTR command. To retrieve this program, type

```
FTR  <sysgrp>common>class>boxplot.run  fts_depot>boxplot.run  -ss RVARES
```

and the program will be copied to the fts_depot on your Prime. The Fortran source code for this program (`boxplot.f77`) can be found in the same directory. Instructions for use of the program can be found at the beginning of the source code, and in a file called `**read.me.first`, also in the common>class directory. Substitute those file names for `boxplot.run` in the FTR command above to retrieve those files to your fts_depot, or see your site administrator for help. This program produces a Telagraf file, consisting of side-by side boxplots and appropriate labels. Figure 2 was constructed using it.

This boxplot program has also been adapted for and implemented in the QWGRAPH component of the QWDATA module of the National Water Information System (NWIS) on the Primes. Contact Kerry Garcia (KTGARCIA,
FTS 702-887-7659) in the Nevada District office for more detail on producing boxplots under NWIS. The program is also available in the DATAGRAF software now on the Primes in most District offices. Contact Jim Schornick (JSCHORNICK, FTS 959-6867) for more information on DATAGRAF.

Figure 2. TDS ALONG THE COLORADO R.



## Boxplots for Censored Data

Data sets are frequently encountered whose values include some observations known only to be below (or above) a limit or threshold. Such data sets are called "censored data" in the statistical literature. Examples in hydrology include chemical concentrations below a detection limit, bacterial colonies too dense to count, depths to water below the bottom of well screens, or flood peaks known only to be lower (or higher) than some known stage. Boxplots can effectively display such data.

Suppose the Figure 2 data were censored with an artificial detection limit, so that some values were recorded only as "less than 600 mg/L". To construct a boxplot for this data (Figure 3), first draw a line across the graph at the value of the detection limit. Next, all lines below this value are erased from the graph. Construction of boxplots for censored data requires that all values below the detection limit are represented by some value less than (not equal to) the detection limit. The actual value is not important, and could be 0, one-half the detection limit, etc.

Figure 3 was constructed using a modification of the boxplot.f77 program named bpcens.f77 (the runfile is in bpcens.run). This program is available by substituting "bpcens" for "boxplot" in the ftr command:

```
FTR  <sysgrp>common>class>bpcens.run  fts_depot>bpcens.run  -ss RVARES
```

If less than 25 percent of the data are below the detection limit, the above procedure will affect at most only the lower whisker (as in the Hoover Dam through Morelos Dam boxes of Figure 3). If between 25 and 75 percent are below the detection limit, the box will be partially hidden below the threshold (as in the CO-UT Line and Cisco boxes). If more than 75 percent of the data are below the detection limit, part of the upper whisker and outside values will be visible above the threshold, as in the Lees Ferry box. In each case, these boxplots accurately and fairly illustrate both the distribution of data above the detection limit, and the percentage of data below the detection limit.

A second alternative for boxplots of censored data is to estimate the percentiles falling below the threshold, and drawing dashed portions of the box below the threshold using these estimates (Figure 4). Helsel and Cohn (1988) have compared methods for estimating these percentiles, and either the lognormal probability plot method or maximum likelihood method should be used. Substituting one-half the detection limit, etc. for each observation will **not** produce as good of an estimate -- see Helsel and Cohn (1988) for more detail.

When multiple thresholds occur, such as detection limits which have changed over time or between laboratories, a solid line can be drawn across the plot at the highest threshold. Portions of the boxes above the highest threshold will be correct as long as each censored observation is assigned some value below its threshold. Quartiles falling below the highest threshold should be determined by using the methods recommended by Helsel and Cohn (1988). All lines below the highest threshold are estimates, and should be drawn as dashed lines on the plot.

Figure 3. TDS BOXPLOTS WITH DETECTION LIMIT

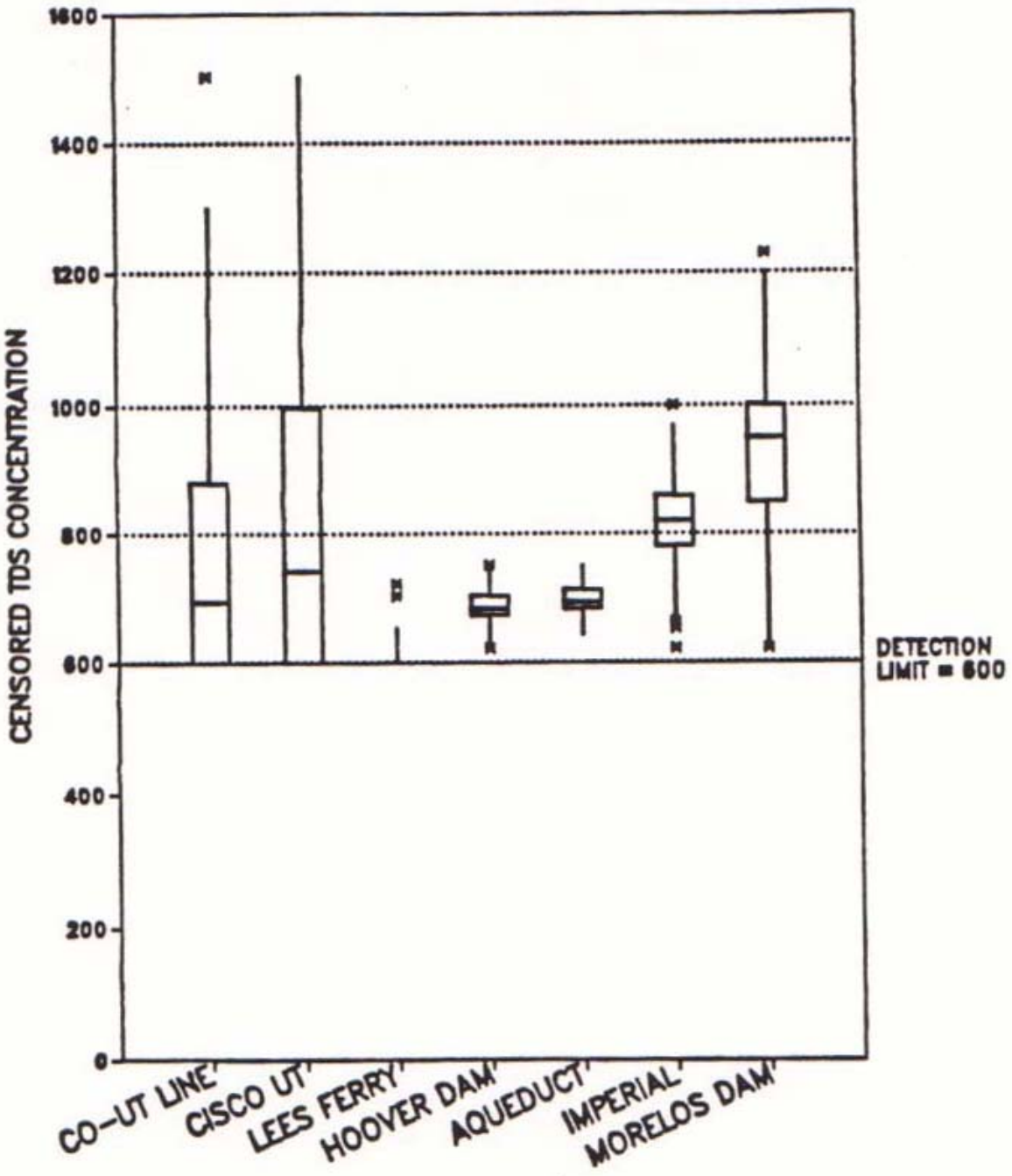
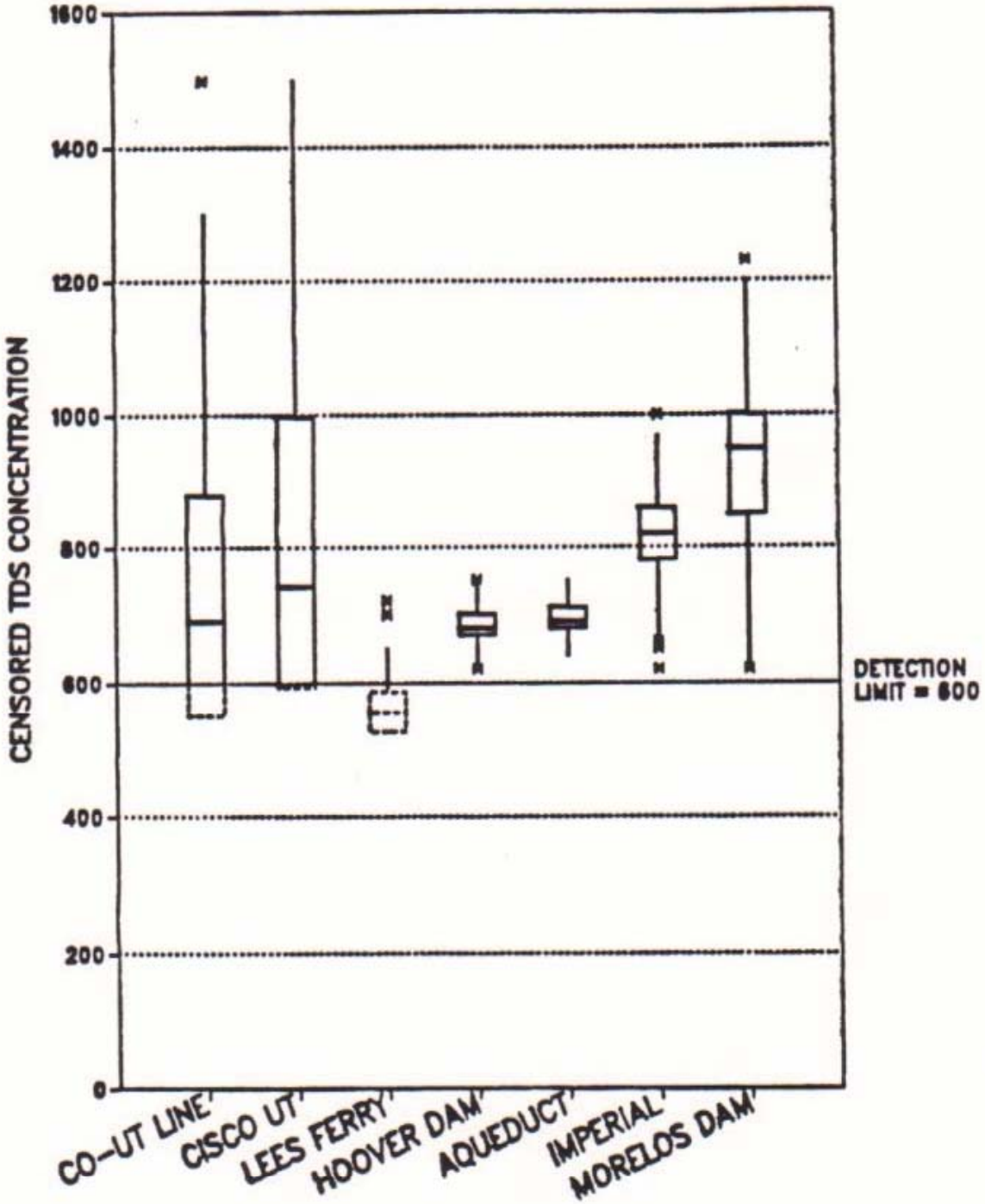


Figure 4. TDS BOXPLOTS USING ESTIMATES



## Appendix: Details of Boxplot Construction

The upper and lower limits of the box are defined as either quartiles or fourths. These definitions are clarified below. Then the influence of each definition on the position of the whiskers is demonstrated. Definitions used by commercial software packages are listed, including one non-conventional form called a "box graph".

Quartiles

Quartiles are the 25th, 50th and 75th percentiles of a data set. The 50th percentile ($p_{.50}$) is also called the median. The 75th percentile, or upper quartile ($p_{.75}$), is a value which exceeds no more than 75 percent of the data and is exceeded by no more than 25 percent of the data. The lower quartile ($p_{.25}$) is a value which exceeds no more than 25 percent of the data and is exceeded by no more than 75 percent. Consider a data set $X_i$, i=1,...n. Computation of percentiles follows the equation

$$p_j = X_{(n+1)\bullet j}$$

where n is the sample size of $X_i$,

j is the fraction of data less than or equal to the percentile value (for the 3 quartiles, j= .25, .50, and .75), and

non-integer values of $(n+1)\bullet j$ imply linear interpolation between adjacent values of X.

Computation of quartiles for two small example data sets is illustrated in Table 1.

Fourths

Tukey (1977) used values for the ends of the box which, along with the median, divided the data into four equal parts. These "fourths" (also called "hinges") are defined as:

Lower fourth $f_L$ = median of all observations less than or equal to the sample median.

Upper fourth $f_U$ = median of all observations equal to or greater than the overall sample median.

They may also be defined as:

Lower fourth $f_L = X_L$, where $L = \dfrac{\text{integer}\,[\,(n+3)/2\,]}{2}$, and

Upper fourth $f_U = X_U$, where $U = (n+1) - L$.

where "integer [ ]" is the integer portion of the number in brackets. For example, integer [ 5.7 ] = 5. Again, non-integer values of L and U imply interpolation. With fourths, however, this will always be halfway between adjacent data points. Therefore, fourths are always either data values themselves, or averages of two data points, and so

are easier to compute by hand than are percentiles. Fourths will generally be similar to quartiles for large (n>30) sample sizes. For smaller data sets, differences will be more apparent. For example, when n=12 the lower fourth is halfway between the 3rd and 4th data points, while the lower quartile is one-quarter of the way between the two points (see Table 1) . Both measures split the data into one-fourth below and three fourths above their value. Either are acceptable for use in boxplots.

Table 1

For sample size n=11, and data $X_i$, i=1,...n equal to:
2 3 5 45 46 47 48 50 90 151 208

$p_{.25}$ = lower quartile = $X_{(n+1)\bullet.25} = X_3 = 5.$
$p_{.75}$ = upper quartile = $X_{(n+1)\bullet.75} = X_9 = 90.$

$p_{.50}$ = median = $X_{(n+1)\bullet.50} = X_6 = 47.$

$f_l$ = lower fourth = median [2 3 5 45 46 47] = 25.
$f_u$ = upper fourth = median [47 48 50 90 151 208] = 70.

For sample size n=12, and data $X_i$, i=1,...n equal to:
2 3 5 45 46 47 48 49 50 90 151 208

$p_{.25}$ = lower quartile = $X_{(n+1)\bullet.25} = X_{3.25} = X_3 + 0.25\bullet(X_4 - X_3)$ = 15.
$p_{.75}$ = upper quartile = $X_{(n+1)\bullet.75} = X_{9.75} = X_9 + 0.75\bullet(X_{10} - X_9)$ = 80.

$p_{.50}$ = median = $X_{(n+1)\bullet.50} = X_{6.5} = X_6 + 0.50\bullet(X_7 - X_6)$ = 47.5.

$f_l$ = lower fourth = median [2 3 5 45 46 47] = 25.
$f_u$ = upper fourth = median [48 49 50 90 151 208] = 70.

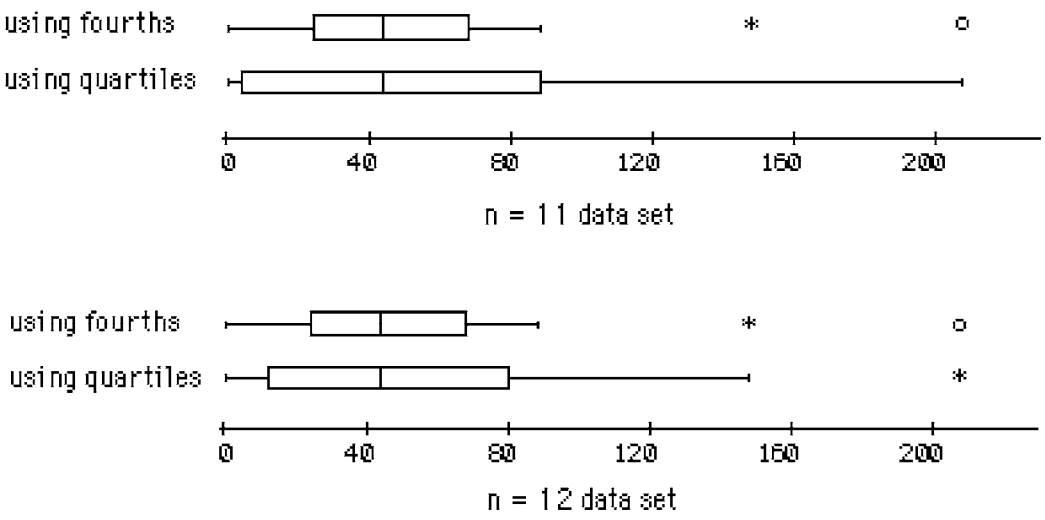



Figure 5. Boxplots for the Table 1 data

Figure 5 shows standard boxplots for the Table 1 data using both percentiles and fourths. Data in Table 1 were designed to maximize differences between the two measures. Real data, and larger sample sizes, will evidence much smaller differences. Note that the definitions of the box boundaries directly affect whisker lengths, and also determines which data are plotted as "outside" values.

It would be ideal if everyone used the same conventions for drawing boxplots. However, that has not happened. Minitab, PSTAT, Systat, DataDesk, and the USGS code listed above follow the original use of fourths, while SAS and S use the quartile values. Those who stick to the original definition prefer fourths; those who want box boundaries to agree with tabled percentiles use quartiles. The Table 1 data can be used to determine which convention is used for other software that produce boxplots.

Non-conventional definitions

SPSS and Statgraphics use another (non-conventional) value for the box boundaries (Frigge and others, 1989). They use the next highest data value for the lower box boundary whenever n/4 is not an integer. This avoids all interpolation. Note that n, not n+1, is used.

StatView (512+ and II) uses a percentile-type boxplot similar to the truncated boxplot, except that the upper and lower 10 percent of data are plotted as individual points. The weakness of using percentiles for determining whisker length is that 10 percent of the data will always be plotted individually at each end of the plot, and so the plot is less effective for defining and emphasizing unusual values. Also important is that StatView uses yet another non-conventional definition for the box boundaries, $X_{(n+2)\bullet j}$, in calculating the quartiles. This non-conventional boxplot was called a "box graph" by Cleveland (1985).

Therefore SPSS, Statgraphics and especially StatView will produce boxes differing from conventional boxplots, particularly for small data sets.

---

Figures 1 and 5 were produced with Minitab, with arrows and some labels added using Superpaint on a Macintosh Plus.

Figures 2 through 4 were produced using the USGS fortran programs cited in this memo. Program output is as a Telagraf file.

## Summary

1. The primary use of boxplots is for comparison between data sets. Side-by-side boxplots will illustrate the differences in center values, variability around those values, and symmetry of the data sets. Thus boxplots will indicate whether parametric statistical tests are appropriate for determining significant differences between groups. They effectively illustrate the test results, and so are quite useful as figures in the final report. Differences between groups are easier to see with boxplots than with either side-by-side or overlapping histograms. Boxplots better represent data than 'dot and line' plots of means and standard deviations. They are a useful indicator of unusual values, which should be checked for errors but which are often valid data points not to be discarded.

2. Boxplots are available on commercial statistical software packages, and on the Primes. When evaluating software, the ability to produce side-by-side boxplots is important. SAS, for example, does not presently do this.

3. All forms of boxplots give visual clues of center, spread and skewness. Comparisons between boxes should certainly be done using the same method for constructing each box. As data sets become larger, differences in methods of constructing the box become smaller. Do not assume software uses a conventional method of construction. SPSS, Statgraphics and StatView do not.

4. The most common method of construction uses fourths. These will not agree with tabled quartiles for small data sets. If this is of concern, adopt SAS's use of actual quartiles for box boundaries.

5. "Box graphs" with whiskers drawn to the 10th and 90th percentiles give few visual clues of 'outliers'. Ten percent of the data at either end will always be plotted individually, whether or not any or all are considered unusual.

6. Boxplots can be easily modified to illustrate censored data sets, such as data which include values below one or more detection limits. A line should be drawn across the plot at the threshold value, and either all lines erased below this threshold, or missing percentiles estimated and outlines of the box dashed in.

## References

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A., 1983, Graphical Methods for Data Analysis: Duxbury Press, Boston, 395 p.

Chen, H. and Druliner A. D., 1987, Nonpoint-source agricultural chemicals in ground water in Nebraska--Preliminary results for six areas of the High Plains Aquifer: U.S. Geological Survey Water Resources Investigations Report 86-4338, 68 p.

Cleveland, W. S., 1985, The Elements of Graphing Data: Wadsworth Books, Monterey, California, 323 p.

Frigge, M., Hoaglin, D. C., and Iglewicz, B., 1989, Some implementations of the boxplot: American Statistician, v. 43, n. 1, p. 50-54.

Grady, S. J ., and Weaver, M. F., 1987, Preliminary appraisal of the effects of land use on water quality in stratified-drift aquifers in Connecticut: U.S. Geological Survey Water Resources Investigations Report 87-4005, 41 p.

Helsel, D. R. and Cohn, T. A., 1988, Estimation of descriptive statistics for multiply censored water quality data: Water Resources Research, v. 24, n. 12, p. 1997-2004.

Hren, J., Wilson, K. S., and Helsel, D. R., 1984, A statistical approach to evaluate the relation of coal mining, land reclamation and surface-water quality in Ohio: U.S. Geological Water Resources Investigations Report 84-4117, 325 p.

McGill, R., Tukey, J. W., and Larsen, W. A., 1978, Variations of box plots: American Statistician, v. 32, n. 1, p. 12-16.

Moody, D. W., Carr, J., Chase, E. B., and Paulson, R. W., eds., 1988, *in* National Water Summary 1986: U.S. Geological Survey Water-Supply Paper 2325, 560 p.

Schertz, T. L. and Hirsch, R. M., 1985, Trend analysis of weekly acid rain data-- 1978-83: U.S. Geological Water Resources Investigations Report 85-4211, 64 p.

Tukey, J. W., 1977, Exploratory Data Analysis: Addison-Wesley, Reading, Mass., 688 p.

Velleman, P. F. and Hoaglin, D. C., 1981, Applications, Basics, and Computing of Exploratory Data Analysis: Duxbury Press, Boston, 354 p.