Gordon Conference Grants for Beginning Researchers

Through a generous donation from the Statistics Division the 1991 Gordon Research Conference in Chemistry and Chemical Engineering was able to provide grants to pay for the registration, lodging, and food for six beginning researchers. The winners of this year's competition were

*Benjamin M. Adams, Assistant Professor, Department of Math Science and Statistics, University of Alabama

*Nadine E. Chase, Graduate Student, Department of Civil Engineering and Operations Research, Princeton University

*Dennis Lin, Assistant Professor, Department of Statistics, University of Tennessee

*Frederik Lindgren, Graduate Student, Department of Organic Chemistry, University of Umea (Sweden)

*Cynthia Lowry, Assistant Professor, Department of Decision Sciences, Texas Christian University

*Derrick Rollins, Assistant Professor, Departments of Chemical Engineering and Statistics, Iowa State University

MINI PAPER

SHAPE-FINDER BOX PLOTS

Muhammad Aslam
Department of Statistics
University of Karachi, Pakistan

Gerald B. Heyes
Quality Assurance Department
Rank Video Services America
Northbrook, Illinois

Anwer Khurshid
Department of Statistics
University of Karachi, Pakistan

ABSTRACT

Since it's introduction by Tukey (1977), Exploratory Data Analysis (EDA) has attracted a great deal of attention. The authors propose a new modification to the box plot indicating degree of "peakedness" via an estimate of Kurtosis. This estimate becomes a single line added to the interior of the box. The position of the line relative to the theoretical normal (location of the whiskers at the midpoint), represents degree of Kurtosis.

KEYWORDS:

Exploratory Data Analysis, Box Plot, Shape-Finder Box Plot, Kurtosis, Percentiles.

INTRODUCTION

Among recent trends in statistics, Exploratory Data Analysis (EDA) has attracted a great deal of attention. Tukey (1977) first proposed the use of EDA for visual data displays, and for preliminary evaluation of data prior to subjecting them to formal analyses. In particular, both Stem-and-Leaf plots and Box-and-Whisker plots display the form of a data set, ie; whether it is skewed, symmetric, unimodal etc. These plots can also help detect clusters, and outliers. For construction and review see Heyes (1985).

The literature records extensive studies on variations of Stem-and-Leaf and Box plots (Ahmed and Aslam 1988, Becketti and Gould 1987, Heyes 1988, Hunter 1988, McGill, Tukey and Larsen 1987). A new box plot modification is proposed here which provides additional information about the shape or "peakedness" of a data set via a graphic indicator of Kurtosis. The authors suggest the name Shape-Finder Box Plot for this modification.

THE SHAPE-FINDER BOX PLOT

A modification of the conventional box-and-whisker plot, the shape-finder box plot is constructed by simply drawing an additional line within the box. The location of this line relative to the center of the box width where the whisker is normally connected, provides a convenient visual indicator of kurtosis.

ESTIMATING KURTOSIS

While there is considerable disagreement in the literature regarding the definition, meaning and interpretation of Kurtosis, the present authors prefer the simple definition of Moors. Moors (1986) argues that kurtosis is a measure of dispersion in the middle of the distribution and concentration of observations in the tail of a distribution. The concentration in the middle of the distribution may be high or low, consequently the tails are sharp or heavy; hence the distri-

1992 FALL TECHNICAL CONFERENCE CALL FOR PAPERS

A Call for Papers has been issued for the 36th Annual Fall Technical Conference to be held October 8-9, 1992 in Philadelphia, PA. The conference is co-sponsored by the Chemical and Process Industries Division and Statistics Division of the American Society for Quality Control, and the Section on Physical and Engineering Sciences of the American Statistical Association. The theme for the 1992 conference is "Competitiveness Through Continuous Improvement: A Nation Waking Up".

Applied or expository papers are needed for Statistics, Quality Control, and Tutorial/Case Study sessions. Statistics papers should be at or below the level of Technometrics, while Quality Control and Tutorial/Case Study papers should be at or below the level of Journal of Quality Technology.Please send the title, one-page abstract, and one-page outline of your paper by January 24, 1992 to Greg Piepel, Battelle-Northwest, P.O. Box 999, Richland, WA 99352. For more information contact any of the program chairs: Greg Piepel of the Section on Physical and Engineering Sciences (509-375-6911), Kymberly Hockman of the Chemical and Process Industries Division (302-999-6634), or Ralph St. John of the Statistics Division (419-372-8098).

bution may be leptokurtic, mesokurtic or platykurtic respectively. Further, the authors have chosen a simple measure of kurtosis which is consistent with other boxplot measures. This measure, K is called the percentile coefficient of kurtosis and is given by:

$K=((Q3-Q1)/2)/(P_{90}-P_{10})$ where Q_1 and Q_3 are lower and upper quartiles, or by

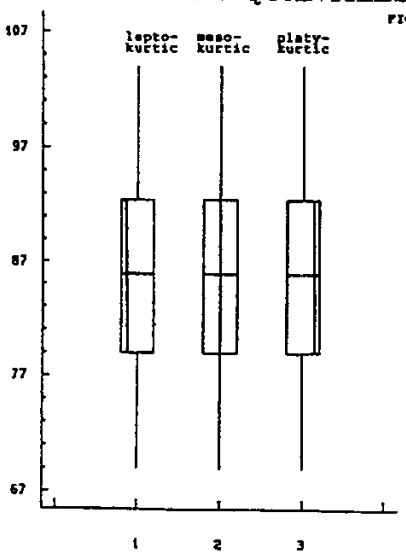
$K=(HS/2)/(P_{90}-P_{10})$

where HS is the Hinge Spread to be discussed later.

P_{10} and P_{90} are the 10th and 90th percentiles respectively of the ordered data. A standard value of K=26.3% is obtained for the normal, mesokurtic distribution. Leptokurtic and platykurtic distributions obtain K values less than 26.3% or more than 26.3% respectively, as shown in Figure 1.

COMPUTING QUANTILES

FIGURE 1



Box Plot Degrees of Kurtosis

To obtain the necessary hinges, percentiles and median the authors recommend using a stem-and-leaf plot as described in example 1.

EXAMPLE 1

Table 1 shows pulse rate measurements from 45 students.

79	88	97	94	79
94	87	75	81	80
91	88	74	79	80
90	88	68	81	81
87	91	70	80	82
92	92	80	82	83
91	90	64	70	84
92	90	67	79	84
94	90	68	83	90

TABLE 1

Table 2 displays the pulse rate data in stem-and-leaf format using double stems (ie: *60 =60 thru 64, while .60 =65 thru 69). Using the stem-and-leaf plot we find the values corresponding to the "depth" of the measures of interest.

TABLE 2

Stem	Ordered Leaves	No. Leaves	Cum. Total
*60	4	1	1
.60	7 8 8	3	4
*70	0 4	2	6
.70	5 8 9 9 9 9	6	12
*80	0 0 0 0 1 1 1 2 2 3 3 4 4	13	25
.80	7 7 8 8 8	5	30
*90	0 0 0 0 0 1 1 1 2 2 2 4 4 4	14	44
.90	7	1	45

The "depth" of a measure is given by formulas and computed as below. See Freund and Perles 1987 for the d(P) formula.

Sample size N=45

d(M)=Depth of Median =(N+1)/2 =46/2 =23

d(H)=Depth of Hinges =(d(M)+1)/2 =24/2 =12

d(P)=Depth of Percentiles =1+((N-1)/10) =5.4

The values corresponding to the depth of the measures are obtained by counting inward from the largest (or smallest) values in the ordered data. For instance the lower Hinge is located by counting in 12 values from and including, the smallest observation. The 12th value of 79 is easily located from the stem-and-leaf plot in table 2. Table 3 summarizes the depths and corresponding values for the measures in example 1.

TABLE 3

Measure	Depth	Value	Plot Feature
Median (M)	23	83	Median Line
Hinges (H)	12	79,90	Box Length
Min, Max	1	64,97	Whisker Ends
Percentiles (P)	5.4	71.6,92	

BOX PLOT CONSTRUCTION

First draw a simple box plot with length equal to the interquartile range, IQR =(Q3-Q1) or the Hinge Spread HS. The whiskers extend to the highest and lowest values, and are located outside the box attached to the center. A median line is added inside the box as in Figure 2. Figure 3 shows a box plot of the student pulse rate data, using hing rather than quartiles.

THE SHAPE-FINDER BOX PLOT MODIFICATION

For the shape finder box plot we will assume a standard K value of 26.3% for the location of the whiskers at the ends of the box. Using Moors' estimate, for example 1 kurtosis is computed as:

K=((UpperH-LowerH) /2)/(P_{90}-P_{10})

K=(90-79)/2 /(92-71.6)

K=(11/2) /20.4 = 5.5/20.4 =26.96%

Having calculated K, it may now be used to create a shape-finder box plot for the student pulse rates. Scaling the end of the box plot enables us to indicate the "percentile coefficient of Kurtosis" on the box, by comparing it to the theoretical normal distribution at the whiskers where K=26.3%.

The method of scaling to indicate percentiles on a box plot is derived from Cleveland (1985). First, assign a the standard value of 26.3% to the location on the x-axis corresponding to the lower whisker location. Next assign an x value of 0 to the side of the box to the left of the whisker Finally the value of 52.6% i assigned to the x location corresponding to the right side of the

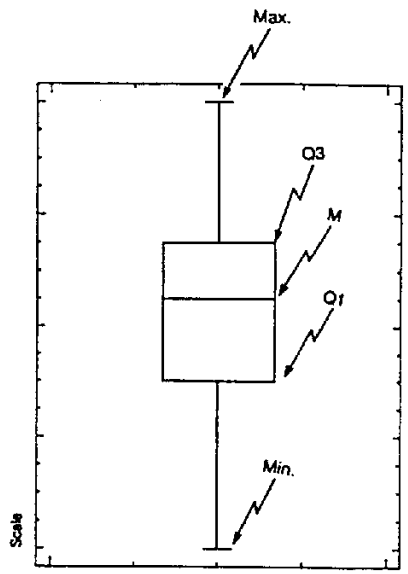


FIGURE 2

box. By this method each side of the box is 26.3% units from the central whisker. For computed values of K greater than 52.6%, simply continue the scale upwards in major increments of 26.3% truncated at the ending value of 100%.

From table 3, the subsequent calculation of K, and the x- axis scaling, construction of the shape-finder box plot is illustrated in figure 4. In this example, other box plot features (fences, confidence envelopes etc.), have been omitted to draw attention to the location of the K value line.

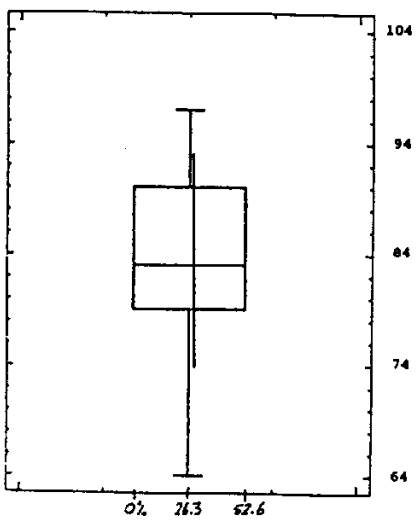


FIGURE 4
SHAPE-FINDER MODIFICATION

Figure 4 displays the following obvious properties:

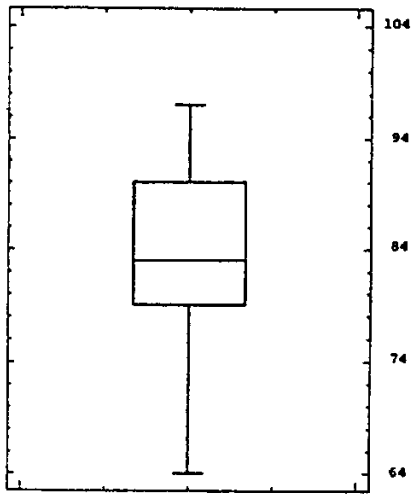


FIGURE 3
STUDENT PULSE RATES

I) The Median value is about 83
- Location of internal median line relative to X-axis scale
- Also known from Table 3

II) There is a slight positive skew
- Location of median line relative to the hinges (box ends)
- Relative lengths of the two whiskers- Confirmed via computed coefficient of skewness -0.583

III) There are no outliers in the sample data
- No values exceeding computed inner/outer fences

IV) The data shape appears to be almost mesokurtic
- Location of K value line relative to whiskers
- shape of the stem-and-leaf plot

REFERENCES

Ahmed, E. and M. Aslam 1988. Extended Digidot Plot. unpublished manuscript.

Becketti, S. and W. Gould 1983. "Rangefinder Box Plots: A Note."The American Statistician, Vol 41, p. 149.

Cleveland, W.S. 1985. The Elements of Graphing Data. Wadsworth Pub. Monterey, CA. pp. 127-144.

Freund, J.E. and B.M. Perles, 1987. "A New Look at Quartiles of Ungrouped Data." The American Statistician, Vol 41, pp. 200-202.

Heyes, G.B. 1985. "The Box Plot." Quality Progress, Vol XVIII,No 12, pp. 12-17.

Heyes, G.B. 1988. "The Ghost Box Plot." ASQC Stat. Div. Newsletter, Vol 8, No 12, pp. 10-11.

Hunter, J.S. 1988. "The Digidot Plot." The American Statistician, Vol 42, No. 1, p. 54.

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

Moors, J.J.A. 1986. "The Meaning of Kurtosis: Darlington Reexamined." The American Statistician, Vol 40, pp. 283-284.

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

THE AUTHORS

Gerald B. Heyes is a senior member of the American Society for Quality Control, a C.Q.E., and past member of the American Chemical Society and the American Management Association. Experienced in chemical, aerospace and electronics industries, he is Senior Q. A. Engineer for Rank Video Service America, in Northbrook, IL. Mr. Heyes has published numerous articles dealing with the use of statistics in industry and taught classes in SPC and process troubleshooting. "Jed" earned a BA (honors) in Biology from Blackburn College in Carlinville, IL.

Muhammad Aslam is a lecturer in the Department of Statistics, University of Karachi, Pakistan. He holds a Masters degree in Statistics from the University of Karachi.

Anwer Khurshid is an Assistant Professor in the Department of Statistics, University of Karachi and a member of the International Association for Official Statistics, ISI, Netherlands.

ADDENDUM

The main idea, that of modifying a box plot with Moor's estimated K value as a line, was presented to Jed Heyes some time ago by co-authors Aslam and Khurshid. Recognizing that the central theme may have some merit, Heyes responded to their request for assistance in developing it thus far. Jed would be delighted to hear from readers who have comments or suggestions for improvement. Some already made include: (1) Determine the operating characteristics of this K statistic relative to the classic B2 measure. (2) Find the optimum scaling for the ends of the box as opposed to the most convenient and determine it's sensitivity to K values. (3) Alternatively, scale the ends of the box based on computed K values for distributions at the extremes of Kurtosis, ie; the most common leptokurtic and platykurtic. (What are these?) (4) Standardize this unusual K value of 26.3% by dividing it by 26.3% thus the standard normal would have a standardized K value of 1.0. (5) How useful is this as a tool?

Editor