Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 47 (2004) 165–174
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A simple more general boxplot method for identifying outliers

Neil C. Schwertman[a,*], Margaret Ann Owens[a], Robiah Adnan[b]

[a] Department of Mathematics & Statistics, California State University, Chico, Chico, CA 95926-0525, USA
[b] Department of Mathematics, Faculty of Science, University of Technology Malaysia, KG791, Johor Bahru 80990, Malaysia

Accepted 1 September 2003

Abstract

The boxplot method (Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1977) is a graphically-based method of identifying outliers which is appealing not only in its simplicity but also because it does not use the extreme potential outliers in computing a measure of dispersion. The inner and outer fences are defined in terms of the hinges (or fourths), and therefore are not distorted by a few extreme values. Such distortion could lead to failing to detect some outliers, a problem known as "masking".

A method for determining the probability associated with any fence or observation is proposed based on the cumulative distribution function of the order statistics. This allows the statistician to easily assess, in a probability sense, the degree to which an observation is dissimilar to the majority of the observations. In addition, an adaptation for approximately normal but somewhat asymmetric distributions is suggested.

Keywords: Hinge; Asymmetric; Fences

1. Introduction

It is well known that outliers, observations that are presumed to come from a different distribution than that for the majority of the data set, can have profound influence on the statistical analysis and can often lead to erroneous conclusions. Because outliers

* Corresponding author.
E-mail address: nschwertman@csuchico.edu (N.C. Schwertman).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2003.10.012

and other extreme values can be very influential in most parametric tests, it is not surprising that the detection and accommodation of outliers have received considerable attention in the literature (see for example: Andrews (1974); Andrews and Pregibon, 1978; Atkinson, 1994; Bacon-Shone and Fung, 1987; Barnett, 1978; Barnett and Lewis, 1995; Brant, 1990; Bhandary, 1992; Gnanadesikan and Kettenring, 1972; Hampel, 1985; Hawkins, 1980; Penny, 1996 and Tukey, 1977). The extensive literature on the subject of outliers attests to its relevance as a major concern in the statistical analysis of data.

One simple way commonly employed to identify outliers is based on the concept of the boxplot and involves the use of "inner fences" and "outer fences." This method, suggested by Tukey (1977), has come into common usage, is often included in texts (see for example, Milton, 1999 pp. 55–58), and has been studied extensively (see for example, Hoaglin et al., 1986; Carling, 2000; Beckman and Cook, 1983; Frigge et al., 1989). This graphically-based method for identifying outliers is especially appealing not only in its simplicity but, more importantly, because it does not use the extreme potential outliers which can distort the computing of a measure of spread and lessen the sensitivity to outliers. The fences procedure uses the estimated interquartile range often referred to as the H-spread, which is the difference between values of the hinges, i.e., sample third and first quartiles. Specifically, the inner fences, f_1 and f_3, and outer fences, F_1 and F_3, are usually defined as

$$f_1 = q_1 - 1.5H\text{-spread},\ f_3 = q_3 + 1.5H\text{-spread},\ F_1 = q_1 - 3H\text{-spread and}$$
$$F_3 = q_3 + 3H\text{-spread}, \tag{1.1}$$

where q_1 and q_3 are the first and third sample quartiles and H-spread $= q_3 - q_1$. Tukey (1977) called observations that fall between the inner and outer fences in each direction "outside" outliners, while those that fall below the outer fence F_1 or above the outer fence F_3 are "far out" outliers.

Inconsistency in defining quartiles complicates the construction of fences. Carling (2000) points out that various authors, such as Cleveland (1985), Freund and Perles (1987), Frigge et al. (1989), Hyndman and Fan (1996), Harrell and Davis (1982), Hoaglin et al. (1983) and Hoaglin and Iglewicz (1987) have proposed a variety of definitions for quartiles. Frigge et al. (1989) list eight definitions of the sample quartiles. With so many definitions it is not surprising that there is variation in determining the criteria for outliers. Carling (2000) compares the outside rate for two of the most common definitions, one by Tukey and the "ideal or machine fourth" recommended by Frigge et al. (1989), concluding the "ideal or machine fourth" is an improvement.

In this paper, the sample quartiles or hinges are approximated by finding the middle point of a set of ordered observations and then finding the approximate quartiles q_1 and q_3 as the middle points of the ordered smaller and larger sets, respectively. See Tukey (1977, pp. 29–38) for more details on the method of finding "hinges".

Many practitioners might find the outer fences too conservative, causing them to overlook many real outliers. In this paper a new simple more general fences method is suggested which allows flexibility in setting the "outside rate", that is, the probability that an observation from a non-contaminated normal population is outside a specified limit or boundary. While the theoretical development assumes both a normal population and a very large sample, this paper also includes a table to accommodate small sample sizes. Many distributions encountered in practice have thicker tails than the normal distribution. For such cases, the adaptation of fences for non-normal and asymmetric distributions is discussed.

2. Establishing a probabilistic basis for the "inner" and "outer" fences

The 1.5 and 3.0 constants commonly used to define the inner and outer fences provide the practitioner only a general sense of how extreme an observation might be without assigning a probability to the degree to which the observations are outliers. Hence, we suggest a simple probabilistic basis for the outside rate criterion or the probability for declaring an uncontaminated observation as an outlier. In the development, it is assumed that the data come from a normal population.

When the samples are sufficiently large, using the method of moments, the difference in sample quartiles q_3-q_1 approximates the difference in population quartiles Q_3-Q_1. That is,

$$q_3 - q_1 \doteq Q_3 - Q_1 = 2(0.67449)\sigma \quad \text{and} \quad \hat{\sigma} = (q_3 - q_1)/1.34898.$$

Carling (2000) points out: "The outside rate,..., has been found in small samples to deviate considerably from its asymptotic counterpart". For the small samples in the simulation study by Hoaglin et al. (1986), the outside rates showed incorrect identification of an outlier at a rate as high as 8.6% for a 1.5 fence and sample size of 5. Clearly an adjustment for sample size is necessary to maintain a reasonable outside rate. Tukey (1977, pp. 632–633) in exhibit 12 provides the appropriate adjustment for small samples.

As an alternative to Tukey (1977) exhibit 12, an adjustment for smaller samples can be obtained by using the expected value of the sample interquartile range, that is, $E(q_3-q_1)=E(q_3)-E(q_1)$.

Harter (1961) investigated the expected values of the normal order statistics, suggesting the use of the Blom (1958) approximation $E(X_i)=\Phi^{-1}((i-\alpha)/(n-2\alpha+1))$ for a sample of size n from a normal population, where $\Phi(x)$ is the cumulant normal function, X_i is the ith order statistic, and $\alpha \approx 0.393$.

To illustrate, suppose $n=35$. Then q_1 and q_3 are the 9th and 27th ordered observations. Using the Blom approximation this is an unbiased estimate of the 24.442 and 75.558 percentiles of the normal population. That is, $E(X_9)=\Phi^{-1}(0.24442)=-0.692155\sigma$ and $E(X_{27})=\Phi^{-1}(0.75558)=0.692155\sigma$ or 0.692155 standard deviations above the mean. Then $E(q_3-q_1)=E(X_{27}-X_9)=1.38431\sigma$ and, using the method of moments, $\hat{\sigma}=(q_3-q_1)/1.38431$.

Observe that for very large samples,

$$E(q_1) = \lim_{\substack{n\to\infty \\ i \to n/4}} \Phi^{-1}\left(\frac{i-\alpha}{n-2\alpha+1}\right) = \Phi^{-1}(0.25) = 0.67449\sigma$$

and similarly

$$E(q_3) = \lim_{\substack{n\to\infty \\ i \to 3n/4}} \Phi^{-1}\left(\frac{i-\alpha}{n-2\alpha+1}\right) = \Phi^{-1}(0.75) = 0.67449\sigma.$$

It follows that $\hat{\sigma} = (q_3 - q_1)/1.34898$. The fences for approximately an α probability of an uncontaminated observation outside one of the fences (outside rate) is

$$f = q_2 \pm \frac{q_3 - q_1}{1.34898} Z_{\alpha/2} \quad \text{where} \quad P(Z > Z_{\alpha/2}) = \alpha/2. \tag{2.1}$$

When the data are not distributed normally it is possible to use *Mathematica* to find the expected values of the ordered observations, and hence the interquartile range directly. *Mathematica* requires the assumed probability distribution of the data as well as the probability density function of the order statistics, i.e.

$$g(y_i) = \frac{n!}{(i-1)!(n-i)!} F(y_i)^{i-1}[1 - F(y_i)]^{n-i} f(y_i), \tag{2.2}$$

where y_i is the ith ordered observation from a random sample size n from a population with assumed probability density function $f(x)$ and $F(y_i) = \int_{-\infty}^{y_i} f(x)\,\mathrm{d}x$.

Using the method of moments, the ordered ith observation is equated to its expected value,

$$E(y_i) = \frac{n!}{(i-1)!(n-i)!} \int_{-\infty}^{\infty} y_i F(y_i)^{i-1}[1 - F(y_i)]^{n-i} f(y_i)\,\mathrm{d}y_i = k_i\sigma,$$

where k_i is the expected number of standard deviations from the mean for the ith ordered observation.

Using *Mathematica* and Eq. (2.2) it is possible to compute the constant k_n which relates $E(q_3 - q_1)$ to the standard deviation for a sample size n, since $\hat{\sigma} = (q_3 - q_1)/k_n$. Table 1 displays the constants k_n for sample sizes 5–100 and a few larger sample sizes for data from a Gaussian distribution.

To illustrate, assume a sample of size n=35. Using Table 1, $\hat{\sigma}$=q_3−q_2/1.38428 which corresponds closely with $\hat{\sigma} = (q_3 - q_1)/1.38431$ obtained using Blom's approximation.

3. Outliers in asymmetric and non-normal distributions

Carling (2000) suggests that it is advantageous, with asymmetric distributions, to construct fences using a constant multiple of the interquartile range measured from the median, q_2. When the distribution is asymmetric, a single criterion for classification as an outlier may not be appropriate. A specific distance from the median in one direction may be quite extreme but the same distance in the other may not be unusual. This can lead to overlooking outliers in the short tail and conversely identifying too many non-outliers in the long tail. To overcome this effect, Kimber (1990) suggested two different measures of dispersion depending on the direction from the median. Specifically, using the semi-interquartile range instead of using the entire IQR,

Table 1

n	k_n	n	k_n	n	k_n	n	k_n	n	k_n	n	k_n
5	1.65798	22	1.33333	39	1.38071	56	1.34361	73	1.36635	90	1.34535
6	1.28351	23	1.4023	40	1.34165	57	1.3713	74	1.34454	91	1.36267
7	1.51475	24	1.33753	41	1.38021	58	1.34329	75	1.36557	92	1.34562
8	1.32505	25	1.40096	42	1.34104	59	1.37004	76	1.34495	93	1.36258
9	1.50427	26	1.33587	43	1.37779	60	1.34394	77	1.36543	94	1.3455
10	1.31212	27	1.39455	44	1.34226	61	1.36981	78	1.34478	95	1.3621
11	1.45768	28	1.33894	45	1.37737	62	1.34366	79	1.36474	96	1.34576
12	1.32968	29	1.39355	46	1.34175	63	1.36871	80	1.34514	97	1.36201
13	1.45268	30	1.3377	47	1.37536	64	1.34424	81	1.36461	98	1.34565
14	1.32353	31	1.38876	48	1.34278	65	1.36851	82	1.34499	99	1.36157
15	1.42975	32	1.34004	49	1.37501	66	1.34399	83	1.36398	100	1.34588
16	1.33318	33	1.38799	50	1.34235	67	1.36754	84	1.34532	200	1.34740
17	1.42684	34	1.33909	51	1.37331	68	1.3445	85	1.36387	300	1.34792
18	1.32959	35	1.38428	52	1.34322	69	1.36737	86	1.34517	400	1.34818
19	1.41322	36	1.34092	53	1.37301	70	1.34429	87	1.3633	∞	1.34898
20	1.33568	37	1.38367	54	1.34285	71	1.3665	88	1.34548		
21	1.41132	38	1.34017	55	1.37156	72	1.34474	89	1.36319		

that is, SIQR(lower) $= q_2 - q_1$ and SIQR(upper) $= q_3 - q_2$, where q_1, q_2 and q_3 are the first, second and third sample quartiles, SIQR(lower) can be used to identify outliers that are exceptionally small while SIQR(upper) can be used to identify outliers that are extremely large.

For very large samples and α probability of being beyond the fences, replacing $q_3 - q_1$ in Eq. (2.1) with either $2(q_2 - q_1)$ or $2(q_3 - q_2)$ the lower and upper fences are: $f_1 = q_2 - 1.4826\ Z_{\alpha/2}(q_2 - q_1)$ and $f_3 = q_2 + 1.4826\ Z_{\alpha/2}(q_3 - q_2)$.

Carling (2000), considering general non-normal distributions, suggested a somewhat complicated method for establishing the outside rate for such distributions. In practice the data analyst may encounter data which is nearly normal but to some degree asymmetrical (see Fig. 1). Such thin-tailed asymmetric distributions could be modeled by two half-normal distributions with different variances as shown in Fig. 1. Of course, it is the tails that are of interest because this is where the outliers are located. Hence, for identifying outliers, it is only critical that the two half-normal model be a good approximation in the tails. For such near normal but somewhat asymmetric distributions, the two half-normal model can be used as a basis for a simple alternative to Carling's (2000) more general methodology for establishing fences based on probabilities.

To construct fences for near normal but somewhat asymmetrical distributions using the Semi-interquartile range SIQR and small samples, the lower and upper standard deviation estimates are: $\hat{\sigma}_{\mathrm{L}} = 2(q_2 - q_1)/k_n$ and $\hat{\sigma}_{\mathrm{u}} = 2(q_3 - q_2)/k_n$, respectively. Then the lower and upper fences are

$$q_2 - \frac{2(q_2 - q_1)}{k_n} Z_{\alpha/2} \quad \text{and} \quad q_2 + \frac{2(q_3 - q_2)}{k_n} Z_{\alpha/2}, \tag{3.1}$$

respectively.

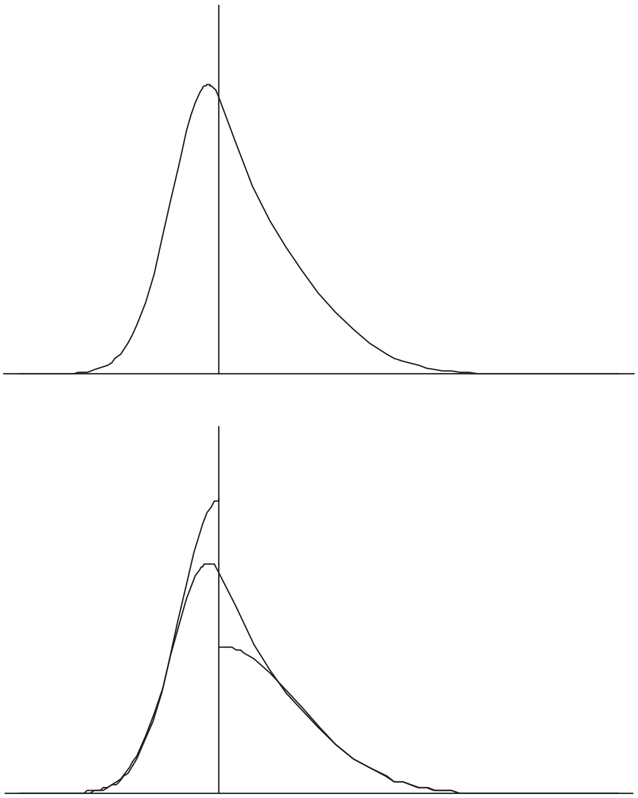

Fig. 1.

To construct fences for a general distribution with outside rate α , the general form of the fences is set equal to $F^{-1}(\alpha/2)$ and $F^{-1}(1-\alpha/2)$. That is $q_1-c_1(q_2-q_1)=F^{-1}(\alpha/2)$ and $q_3+c_2(q_3-q_2)=F^{-1}(1-\alpha/2)$. Substituting the expected values of the qs and solving for c_1 and c_2,

$$c_1=\frac{E(q_1)-F^{-1}(\alpha/2)}{E(q_2-q_1)} \quad \text{and} \quad c_2=\frac{F^{-1}(1-\alpha/2)-E(q_3)}{E(q_3-q_2)}$$

the lower fence is

$$q_1-\frac{E(q_1)-F^{-1}(\alpha/2)}{E(q_2-q_1)}(q_2-q_1)$$

and the upper fence is

$$q_3+\frac{F^{-1}(1-\alpha/2)-E(q_3)}{E(q_3-q_2)}(q_3-q_2).$$

Table 2
Number outside fence (10,000 simulations)

Sample Size	Observation	Contaminated $Y_{(1)}$		No Contamination	
		Usual procedure $f_1 = q_1 - 1.5(q_3 - q_1)$	New procedure $f_1 = q_1 - 3(q_2 - q_1)$	Usual procedure $f_1 = q_1 - 1.5(q_3 - q_1)$	New procedure $f_1 = q_1 - 3(q_2 - q_1)$
30	$Y_{(1)}$	1628	8204	18	1357
70	$Y_{(1)}$	641	8950	1	2037
		Usual procedure $f_3 = q_3 + 1.5(q_3 - q_1)$	New procedure $f_3 = q_3 + 3(q_3 - q_2)$	Usual procedure $f_3 = q_3 + 1.5(q_3 - q_1)$	New procedure $f_3 = q_3 + 3(q_3 - q_2)$
30	$Y_{(30)}$	6024	3184	5799	2954
70	$Y_{(70)}$	8358	3947	8249	3797

For many distributions the computation of the expected value of the interquartile and semi-interquartile ranges as well as $F^{-1}(\alpha/2)$ and $F^{-1}(1-\alpha/2)$ will require numerical integration. If the skewness and kurtosis factors are known, or easy to compute, Carling's (2000) method of constructing fences for a specified outside rate is much easier to use. For normal or near normal distributions, however, Eq. (3.1) and Table 1 are easier to use.

To compare using $q_2 - q_1$ and $q_3 - q_2$ to detect small and large outliers respectively to using the usual $q_3 - q_1$, a simulation study of a skewed distribution was used. The simulated distribution was constructed by generating a standard normal variable z_i. The data xs were

$$x_i = \left\{ \begin{array}{ll} z_i + 5, & z_i < 0 \\ 5(z_i + 1), & z_i \geqslant 0 \end{array} \right\}.$$

The resulting distribution has a variance of 1 to the left and 25 to the right. Ten thousand simulated data sets each of either 30 or 70 observations were simulated. For each data set, the sample quartiles were used to calculate f_1 and f_3 from Eq. (1.1). Data sets both with and without an outlier were simulated. When using an outlier, x_{30} or x_{70} was set to zero. Table 2 displays the results of the simulations.

The simulations followed the anticipated pattern. Using the semi-interquartile range, $(q_2 - q_1)$ to construct the lower fence f_1 provided a fence which was more sensitive for identifying the contaminated observation $y_{(1)}$. Specifically, the procedure using $(q_2 - q_1)$ outperformed the standard procedure using $(q_3 - q_1)$ by correctly identifying the contaminated observation in 82.04% of the simulations compared to only 16.28% for the standard procedure for sample size 30. Similarly, for sample size 70, the correct identification of the contaminated observation rate was 89.5% compared to 6.41% for the standard procedure.

As would be expected, this increased sensitivity for detecting contaminated observations resulted in an increase in the number of observations below the lower fence, f_1, for the uncontaminated data. For the uncontaminated data and sample size 30, the procedure had 13.5% compared to the standard procedure at 0.18% below the lower fence. Similarly for the uncontaminated data and sample size 70 the rate was 20.37% compared to 0.01% for the standard procedure using $(q_3 - q_1)$. Clearly, the greatly increased sensitivity to contaminated observations comes at the expense of increasing the number of uncontaminated observations that are below the lower fence.

For identifying extreme observations in the upper tail for this highly skewed distribution the use of the semi-interquartile range $(q_3 - q_2)$ to construct f_3 based on $(q_3 - q_1)$ was superior to the standard method. Specifically, for sample size 30, the simulated data with the smallest observation contaminated had 60.24% above the upper fence, f_3, for the standard procedure using $(q_3 - q_1)$ and 31.84% using $(q_3 - q_2)$ to construct f_3. Similarly for the completely uncontaminated simulated data 57.99% were beyond the upper fence using $(q_3 - q_1)$ versus 29.54% when constructing f_3 using $(q_3 - q_2)$. For sample size 70, the simulations indicated a similar but more extreme pattern. Specifically, for the simulated data with only the smallest observation contaminated, 83.58% of the observations were beyond f_3 constructed using $(q_3 - q_1)$, while 39.47% where beyond the f_3 fence constructed using $(q_3 - q_2)$. For the completely uncontaminated data the corresponding percentages were 82.49% and 37.97%, respectively. While the proportion beyond the upper fence is very large, using $(q_3 - q_2)$ reduced the proportion by approximately a half in all cases.

For the simulation, using $(q_3 - q_2)$ to construct f_3 greatly improves the probability of identifying a contaminated observation and substantially reduces the proportion of data beyond the fence in the thicker tail.

4. Example

To illustrate the proposed method for detecting outliers, consider the following wood specific gravity data given in Draper and Smith (1966) but as contaminated by Rousseeuw and Leroy (1987):

Ob number	1	2	3	4	5	6	7	8	9	10
Observation	0.534	0.535	0.570	0.450	0.548	0.431	0.481	0.423	0.475	0.486

Ob number	11	12	13	14	15	16	17	18	19	20
Observation	0.554	0.519	0.492	0.517	0.502	0.508	0.520	0.506	0.401	0.568

Observations 4, 6, 8, and 19 are the contaminated data. The sample quartiles are $q_1 = 0.478$, $q_2 = 0.507$ and $q_3 = 0.5345$. The IQR $= q_3 - q_1 = 0.0565$ and, using the usual method of determining fences, the inner fences would be 0.39325 and 0.61925 and the outer fences are 0.3085 and 0.704. None of the contaminated values would have been identified as even mild outliers. Using the procedure proposed in this paper and Table 1,

$$\hat{\sigma}_{\mathrm{L}} = \frac{2(q_2 - q_1)}{k_n} = \frac{2(0.029)}{1.33568} = 0.0434235 \quad \text{and} \quad \hat{\sigma}_{\mathrm{u}} = \frac{2(q_3 - q_2)}{k_n}$$

$$= \frac{2(0.0275)}{1.33568} = 0.0411775.$$

Then the 0.025 and 0.975 fences are $q_2 - 1.96(0.043425)\mathrm{SIQR(lower)} = 0.507 - 0.085 = 0.422$ and $q_2 + 1.96(0.0411775)$ SIQR (upper) $= 0.507 + 0.081 = 0.588$. Similarly computing the 0.05 and 0.95 fences by replacing 1.96 with 1.645, the fences are 0.436 and 0.575 . Contaminated observation 19, was identified at the 0.025 fence, while, in addition, contaminated observations 6 and 8 were identified at the 0.05 fence. The other contaminated observation, number 4, was too close to the non-contaminated data to be detected as an outlier. No observations were larger than the 0.95 fence, so no observations were declared outliers for being too large. Clearly only observation 19 would be identified as a reasonably strong outlier (p value $= 0.0073$).

5. Concluding comments

It is important to identify outliers and extreme values which may have substantial influence on the statistical analysis, leading to distortion and possibly inaccurate conclusions. In this paper, a simple method for constructing boundaries with a specified outside rate (probability of designating a non-contaminated observation an outlier) is suggested which allows the data analyst flexibility is specifying the criteria for outliers.

The usual inner fences f_1 and f_3 given by $f_1 = q_1 - 1.5\mathrm{IQR}$ and $f_3 = q_3 + 1.5\mathrm{IQR}$ for a normal distribution represent approximately 2.70 standard deviations above and below the mean. Thus, the probability of an uncontaminated observation being beyond the inner fences is 0.006976. That is, by chance only 1 in 143 observations would be classified as a “mild” outlier. The usual outer fences F_1 and F_3 given by $F_1 = q_1 - 3\mathrm{IQR}$ and $F_3 = q_3 + 3\mathrm{IQR}$ for the normal distribution represent approximately 4.72 standard deviations from the mean. The probability of an uncontaminated observation being beyond the outer fences is 0.00000235. That is, by chance only 1 in 425,532 observations would be classified as an “extreme” outlier. The outer fences seem to be quite conservative and often may cause outliers to be overlooked. For data from a normal or near normal but somewhat asymmetric distribution, using the value of k_n from Table 1 and an appropriate $Z_{\alpha/2}$to establish the fences

$$f_1 = q_2 - \frac{2(q_2 - q_1)}{k_n} Z_{\alpha/2} \quad \text{and} \quad f_3 = q_2 + \frac{2(q_3 - q_2)}{k_n} Z_{\alpha/2}$$

allows considerable flexibility in setting the outside rate. This flexibility is not characteristic of the usual method of determining fences. For general non-normal distributions, a procedure for constructing fences at a specified probability is proposed. The procedure by Carling (2000), however, seems to be much easier to implement if skewness and kurtosis factors are known.

Acknowledgements

The authors would like to thank the Editor and Associate Editor for their many helpful suggestions which greatly improved this manuscript.

References

Andrews, D.F., 1974. A robust method for multiple linear regression. Technometrics 16, 523–531.

Andrews, D.F., Pregibon, D., 1978. Finding the outliers that matter. J. Roy. Statist. Soc. B 44, 1–36.

Atkinson, A.C., 1994. Fast very robust methods for the detection of multiple outliers. J. Amer. Statist. Assoc. 89, 1329–1339.

Bacon-Shone, J., Fung, W.K., 1987. A new graphical method for detecting single and multiple outliers in univariate and multivariate data. Appl. Statist. 36, 153–162.

Barnett, V., 1978. The study of outliers: purpose and model. Appl. Statist. 27, 242–250.

Barnett, V., Lewis, T., 1995. Outliers in Statistical Data, 3rd Edition. Wiley, Chichester, England.

Beckman, R.J., Cook, R.D., 1983. Outliers........s. Technometrics 25, 119–163.

Bhandary, M., 1992. Detection of the numbers of outliers present in a dataset using an information theoretic criterion. Comm. Statist. Theory Methods 21, 3263–3274.

Blom, G., 1958. Statistical Estimates and Transformed Beta Variable. Wiley, New York.

Brant, R., 1990. Comparing classical and resistant outlier rules. J. Amer. Statist. Assoc. 85, 1083–1090.

Carling, K., 2000. Resistant outlier rules and the non-Gaussian case. Comput. Statist. Data Anal. 33 (3), 249–258.

Cleveland, W.S., 1985. The Elements of Graphing Data. Hobart Press, Summit, NJ.

Draper, N.R., Smith, H., 1966. Applied Regression Analysis, Wiley, New York.

Freund, J.E., Perles, B.M., 1987. A new look at quartiles of ungrouped data. Amer. Statist. 41, 200–203.

Frigge, M., Hoaglin, D.C., Iglewicz, B., 1989. Some implementations of the Boxplot. Amer. Statist. 43, 50–54.

Gnanadesikan, R., Kettenring, J.R., 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28, 81–124.

Hampel, F.R., 1985. The breakdown points of the mean combined with some rejection rules. Technometrics 27, 95–107.

Harrell, F.E., Davis, C.E., 1982. A new distribution-free quantile estimator. Biometrika 69, 635–640.

Harter, H.L., 1961. Expected values of normal order statistics. Biometrika 48, 151–165.

Hawkins, D.M., 1980. Identification of Outliers. Chapman & Hall, London.

Hoaglin, D.C., Iglewicz, B., 1987. Fine-tuning, some resistant rules for outlier labeling. J. Amer. Statist. Assoc. 82, 1147–1149.

Hoaglin, D.C., Iglewicz, B., Tukey, J.W., 1986. Performance of some resistant rules for outlier labeling. J. Amer. Statist. Assoc. 81, 991–999.

Hoaglin, D.C., Mosteller, F., Tukey, J.W., 1983. Understanding Robust and Exploratory Data Analysis. Wiley, New York.

Hyndman, R.B., Fan, Y., 1996. Sample quantiles in statistical packages. Amer. Statist. 50, 361–365.

Kimber, A.C., 1990. Exploratory data analysis for possibly censored data from skewed distribution. Appl. Statist. 39, 21–30.

Milton, J.S., 1999. Statistical Methods in the Biological and Health Sciences. WCB/McGraw-Hill, New York.

Penny, K.I., 1996. Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance. Appl. Statist. 45, 73–81.

Rousseeuw, R.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.

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