Tests to Identify Outliers in Data Series

Francisco Augusto Alcaraz Garcia

1 Introduction

There are several definitions for outliers. One of the more widely accepted interpretations on outliers comes from Barnett and Lewis [1] , which defines outlier as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data". However, the identification of outliers in data sets is far from clear given that suspicious observations may arise from low probability values from the same distribution or perfectly valid extreme values (tails) for example.

One alternative to minimize the effect of outliers is the use of robust statistics, which would solve the dilemma of removing/modifying observations that appear to be suspicious. When robust statistics are not practical for the problem in question, it is important to investigate and record the causes of the possible outliers, removing only the data points clearly identified as outliers.

Situations where the outliers causes are only partially identified require sound judgment and a realistic assessment of the practical implications of retaining outliers. Given that their causes are not clearly determined, they should still be used in the data analysis. Depending on the time and computing power constrains, it is often possible to make an informal assessment of the impact of the outliers by carrying out the analysis with and without the suspicious outliers.

This document shows different techniques to identify suspicious observations that would require further analysis and also tests to determine if some observations are outliers. Nevertheless, it would be dangerous to blindly accept the result of a test or technique without the judgment of an expert given the underlying assumptions of the methods that may be violated by the real data.

2 Z-scores

Z-scores are based on the property of the normal distribution that if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$. Z-scores are a very popular method for labeling outliers and has been implemented in different flavors and packages

as we will see along this document. Z-scores are defined as:

$$Z_{score}(i) = \frac{x_i - \overline{x}}{s}, \quad \text{where} \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1}$$

A common rule considers observations with $|Z_{scores}|$ greater than 3 as outliers, though the criteria may change depending on the data set and the criterion of the decision maker. However, this criterion also has its problems since the maximum absolute value of Z-scores is $(n-1)/\sqrt{n}$ (Shiffler [24]) and it can be possible that none of the outliers Z-scores would be greater than the threshold, specially in small data sets.

3 Modified Z-scores

The problem with the previous Z-score is that the $\overline{x}$ and s can be greatly affected by outliers, and one alternative is to replace them with robust estimators. Thus, we can replace $\overline{x}$ by the sample median ($\tilde{x}$), and s by the MAD (Median of Absolute Deviations about the median):

$$MAD = \text{median}\{|x_i - \tilde{x}|\} \tag{2}$$

Now, the Modified Z-scores are defined as:

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD} \tag{3}$$

where the constant 0.6745 is needed because $\mathbb{E}(MAD) = 0.6745\sigma$ for large n.

Observations will be labeled outliers when $|M_i| > D$. Iglewicz and Hoaglin [13] suggest using $D = 3.5$ relying on a simulation study that calculated the values of D identifying the tabulated proportion of random normal observations as potential outliers.

The suspicious data could be studied further to explore possible explanations for denoting these values as real outliers or not.

This test was implemented in R (see below) with the name of MAD-scores.

4 Boxplot

The main elements for a boxplot are the median, the lower quantile ($Q1$) and the upper quantile ($Q3$). The boxplot contains a central line, usually the median, and extends from $Q1$ to $Q3$. Cutoff points, known as fences, lie $k(Q3 - Q1)$ above the upper quartile and below the lower quartile with $k = 1.5$ frequently. Observations beyond the fences are considered as potential outliers.

Tukey [25] defines the lower fourth as $Q1 = x_f$, the fth ordered observation, where f is computed as:

$$f = \frac{((n+1)/2)+1}{2} \tag{4}$$

If f involves a fraction, $Q1$ is the average of x_f and x_{f+1}. To get $Q3$, we count f observations from the top, i.e., $Q3 = x_{n+1-f}$.

Some other boxplots use cutoff points other than the fences. These cutoffs take the form $Q1 - k(Q3 - Q1)$ and $Q3 + k(Q3 - Q1)$. Depending on the value of k, a different number of potential outliers can be selected. Frigge, Hoaglin and Iglewicz [9] estimated the probability of labeling at least one observation as an outlier in a random normal sample for different values of k, arriving to the conclusion that a value of $k \sim 2$ would give a probability of $5-10\%$ that one or more observations are considered outliers in a boxplot.

5 Adjusted Boxplot

The boxplot discussed before has the limitation that the more skewed the data, the more observations may be detected as outliers. Vanderviere and Hubert [26] introduced an adjusted boxplot taking into account the medcouple (MC), a robust measure of skewness for a skewed distribution.

Given a set of ordered observations, Brys et al. [4] define the MC as:

$$MC = \underset{\substack{x_i \le \tilde{x} \le x_j \\ x_i \neq x_j}}{\text{median}}\, h(x_i, x_j) \tag{5}$$

where the function h is given by:

$$h(x_i, x_j) = \frac{(x_j - \tilde{x}) - (\tilde{x} - x_i)}{x_j - x_i} \tag{6}$$

For the special case $x_i = x_j = \tilde{x}$ the function h is defined differently. Let $m_1 < \ldots < m_q$ denote the indices of the observations which are tied to the median $\tilde{x}$, i.e., $x_{m_l} = \tilde{x}$ for all $l = 1, \ldots, q$. Then:

$$h(x_{m_i}, x_{m_j}) = \begin{cases} -1 & \text{if} \quad i+j-1 < q \\ 0 & \text{if} \quad i+j-1 = q \\ +1 & \text{if} \quad i+j-1 > q \end{cases} \tag{7}$$

According to Brys et al. [3], the interval of the adjusted boxplot is:

$$\begin{aligned} [L, U] &= \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (8) \\ &= [Q1 - 1.5e^{-3.5MC}(Q3 - Q1), Q3 + 1.5e^{4MC}(Q3 - Q1)] \quad \text{if} \quad MC \geq 0 \\ &= [Q1 - 1.5e^{-4MC}(Q3 - Q1), Q3 + 1.5e^{3.5MC}(Q3 - Q1)] \quad \text{if} \quad MC \leq 0 \end{aligned}$$

where L is the lower fence and U is the upper fence of the interval. The observations which fall outside the interval are considered outliers.

The value of the MC ranges between -1 and 1. If $MC = 0$, the data is symmetric and the adjusted boxplot becomes the traditional boxplot for $k = 1.5$. If $MC > 0$ the data has a right skewed distribution, whereas if $MC < 0$, the data has a left skewed distribution.

6 Generalized ESD Procedure

A similar procedure to the Grubbs test below is the Generalized Extreme Studentized Deviate (ESD) to test for up to a prespecified number r outliers. The process is as follows:

1. Compute R_1 from:

$$R_i = max_i\left\{\frac{|x_i - \overline{x}|}{s}\right\} \tag{9}$$

   Then find and remove the observation that maximizes $|x_i - \overline{x}|$

2. Compute R_2 in the same way but with the reduced sample of $n - 1$ observations

3. Continue with the process until $R_1, R_2, \ldots, R_r$ have been computed

4. Using the critical values λ_i at the chosen confidence level α find l, the maximum i such that $R_i > \lambda_i$
   The extreme observations removed at the first l steps are declared as outliers.

For a two-sided outlier problem, the value of λ_i is defined as:

$$\begin{aligned} \lambda_i &= \frac{t_{(p,n-i-1)}(n-i)}{\sqrt{(n-i-1+t^2_{(p,n-i-1)})(n-i+1)}}; \quad i = 1, \ldots, r \\ p &= 1 - \frac{\alpha/2}{n-i+1} \end{aligned} \tag{10}$$

where $t_{(p,d)}$ is the pth percentile of a t distribution with d degrees of freedom. For the one-sided outlier problem we substitute $\alpha/2$ by α in the value of p.

Rosner [22] provides the tabulated values for several α, $n \leq 500$ and $r \leq 10$, and concludes that this approximation is very accurate when $n > 25$.

It is recommended to use this test with a higher number of outliers than expected and when testing for outliers among data coming from a normal distribution.

7 Sample Kurtosis

The sample kurtosis:

$$b_2 = \frac{n \sum_{i=1}^{n} (x_i - \overline{x})^4}{(\sum_{i=1}^{n} (x_i - \overline{x})^2)^2} \tag{11}$$

is used to test for outliers and measure departures from normality.

Initially, b_2 is compared with the critical value for the appropriate n and α. If b_2 exceeds the critical value, then the observation x_j that maximizes $|x_i - \overline{x}|$ is declared an outlier. This outlier is removed and the procedure repeated. If b_2 does not exceed the critical value, then the process stops.

Tables with critical values for $n > 50$ can be found in Pearson and Hartley [18] and for $7 \leq n \leq 50$ in Iglewicz and Hoaglin [13].

Though this is a reasonable procedure to use in practice, it is susceptible to masking when neighboring outliers are present.

8 The Shapiro–Wilk W Test

The Shapiro–Wilk results to test for normality can also be used to test for outliers. Given an ordered sample, the procedure involves:

1. Calculate:

$$b = \sum_{i=1}^{h} a_{n+1-i}(x_{n+1-i} - x_i) \tag{12}$$

   where $h = n/2$ for n even and $(n-1)/2$ for n odd, and the constants a_{n+1-i} can be obtained from different sources.

2. Calculate $D = \sum_{i=1}^{n} (x_i - \overline{x})^2$

3. Compute $W = b^2/D$

4. No outliers are present if $W > C$, where the critical value C is available in a number of sources. Otherwise, consider the most deviant observation from $\overline{x}$ as the outlier. Remove this observation and repeat the process on the reduced sample.

Tables for the critical values and the a_{n+1-i} can be found in Shapiro [23] or Barnett and Lewis [1].

In general, it seems that the generalized ESD test performs better in identifying outliers than the Shapiro-Wilk W test.

9 B1 Statistic for Extreme Deviation

Dixon [6] summarizes several criteria for discovery of one or more outliers of two types entering into samples of observations from a normal population with mean μ and variance σ^2, $N(\mu, \sigma^2)$:

1. One or more observations from $N(\mu + \lambda\sigma, \sigma^2)$
   This is an error in the mean value that is generally referred as "location error".

2. One or more observations from $N(\mu, \lambda^2\sigma^2)$
   This is the occurrence of an error from a population with the same mean but a greater variance than the remainder of the sample and is referred as "scalar error".

The B_1 statistic works on n ordered observations $x_1 < x_2 < \ldots < x_n$ when σ is known or estimated independently. The statistic has the form:

$$B_1 = \frac{x_n - \overline{x}}{\sigma} \quad \text{or} \quad B_1 = \frac{\overline{x} - x_1}{\sigma} \tag{13}$$

and checks if the highest or lowest value in the sample is an outlier.

Grubbs [10] includes in his paper the table of percentile points for B_1 and B_n derived by Pearson and Chandra [17] when σ^2 is the sample variance, and that can be used to test for the rejection-acceptance of the lowest or highest values as outliers. The table provides the values for $3 \leq n \leq 25$ and $\{1\%, 2.5\%, 5\%, 10\%\}$ confidence levels.

If we consider that $B_1^2, B_n^2 \sim \mathcal{X}^2(1)$, the p-value is given as $1 - cdf_{\mathcal{X}^2(1)}(B_1^2, B_n^2)$. Then, the criteria would be that any extreme deviation with p-value $< \alpha$, being α the significant level, is an outlier.

10 Dixon Tests for Outlier

Tests of the Dixon type work with ratios of ranges of parts of an ordered sample that do nor require the knowledge of σ. The different flavors of statistics are [6]:

1. for single outlier x_1 or x_n respectively:

$$r_{10} = \frac{x_2 - x_1}{x_n - x_1} \quad \text{or} \quad r_{10} = \frac{x_n - x_{n-1}}{x_n - x_1} \tag{14}$$

2. For single outlier x_1 avoiding x_n, or x_n avoiding x_1 respectively:

$$r_{11} = \frac{x_2 - x_1}{x_{n-1} - x_1} \quad \text{or} \quad r_{11} = \frac{x_n - x_{n-1}}{x_n - x_2} \tag{15}$$

3. For single outlier x_1 avoiding x_n and x_{n-1}, or x_n avoiding x_1 and x_2 respectively:

$$r_{12} = \frac{x_2 - x_1}{x_{n-2} - x_1} \quad \text{or} \quad r_{12} = \frac{x_n - x_{n-1}}{x_n - x_3} \tag{16}$$

4. For outlier x_1 avoiding x_2, or x_n avoiding x_{n-1} respectively:

$$r_{20} = \frac{x_3 - x_1}{x_n - x_1} \quad \text{or} \quad r_{20} = \frac{x_n - x_{n-2}}{x_n - x_1} \tag{17}$$

5. For outlier x_1 avoiding x_2 and x_n, or x_n avoiding x_{n-1} and x_1 respectively:

$$r_{21} = \frac{x_3 - x_1}{x_{n-1} - x_1} \quad \text{or} \quad r_{21} = \frac{x_n - x_{n-2}}{x_n - x_2} \tag{18}$$

6. For outlier x_1 avoiding x_2, x_n and x_{n-1}, or x_n avoiding x_{n-1}, x_1 and x_2 respectively:

$$r_{22} = \frac{x_3 - x_1}{x_{n-2} - x_1} \quad \text{or} \quad r_{22} = \frac{x_n - x_{n-2}}{x_n - x_3} \tag{19}$$

$r_{11}, r_{12}, r_{20}, r_{21}, r_{22}$ were designed for use in situations where additional outliers may occur and we wish to minimize the effect of these outliers on the investigation of the particular value being tested. According to Walfish [27], situations like these arise because of masking, i.e., when several observations are close together but the group of observations is still outlying from the rest of the data; and it is a common phenomenon specially for bimodal data.

Dixon [7] publishes several tables of critical values (λ_{ij}) for the different statistics r_{ij} for $n \leq 30$ with the criteria for declaring the appropriate $x_{\cdot}$ being an outlier if $r_{ij} > \lambda_{ij}$.

11 Grubbs Test for One or Two Outliers

According to [16], Grubbs test detects one outlier at a time assuming a normal distribution. This outlier is expunged from the dataset and the test is iterated until no outliers are detected. However, multiple iterations change the probabilities of detection, and the test should not be used for sample sizes of six or less since it frequently tags most of the points as outliers.

There are several statistics for the Grubbs test considering an ordered data sample:

1. Test if the minimum or maximum values are outliers

$$G = \frac{\overline{x} - x_1}{s} \quad \text{or} \quad G = \frac{x_n - \overline{x}}{s} \tag{20}$$

where s is the sample standard deviation. This tests looks similar to the B_1 statistic in Section 9 but with the difference that the form of the limiting distribution is different.

This test is also called the Modified Thompson Tau or the maximum normed residual test in other references.

For the two-sided test, the hypothesis of no outliers is rejected if:

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{(\frac{\alpha}{2n}, n-2)}}{n - 2 + t^2_{(\frac{\alpha}{2n}, n-2)}}} \tag{21}$$

with $t_{(\frac{\alpha}{2n}, n-2)}$ denoting the $\frac{\alpha}{2n}$ percentile of a t-distribution with $(n-2)$ degrees of freedom. For one-side tests, we use the $\frac{\alpha}{n}$ percentile.

In the above formulas for the critical regions, the convention used is that t_α is the upper critical value from the t-distribution and $t_{1-\alpha}$ is the lower critical value from the t-distribution.

2. Test for two opposite outliers

$$G = \frac{x_n - x_1}{s} \tag{22}$$

This statistic is referred in Dixon [6] as C_1, and tests simultaneously whether the smallest and largest observations are outlying. David, Hartley and Pearson [5] determine the limiting distribution and Grubbs [11] specifies that the hypothesis of no outliers is rejected if:

$$G > \sqrt{\frac{2(n-1)t^2_{(\frac{\alpha}{n(n-1)}, n-2)}}{n - 2 + t^2_{(\frac{\alpha}{n(n-1)}, n-2)}}} \tag{23}$$

and if x_n is about as far above the sample mean as x_1 is below. If, however, x_n and x_1 are displaced from the mean by different amounts, some further test would have to be made to decide whether to reject as outlying only the lowest value or only the highest value or both the lowest and the highest values.

Nevertheless, Ellison, Barwick and Farrant [8] indicate that the tests are often carried out in turn on the same data set if the single-outlier

test is not significant, to ensure that the single-outlier test is not compromised by a second outlier (as would be detected by the two opposite outlier test). In practice, with a single outlier already identified, one would not normally apply the test for two opposite outliers until the initial single-outlier had been investigated or eliminated.

In spite of this practice, it is important to be aware that using all two (or three) Grubbs tests simultaneously will increase the false-positive rate.

3. Test if the two largest or the two smallest values are outliers

$$\frac{S^2_{n-1,n}}{S^2} = \frac{\sum_{i=1}^{n-2}(x_i - \overline{x}_{n-1,n})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{or} \quad \frac{S^2_{1,2}}{S^2} = \frac{\sum_{i=3}^{n}(x_i - \overline{x}_{1,2})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{24}$$

$$\overline{x}_{n-1,n} = \frac{1}{n-2}\sum_{i=1}^{n-2} x_i \quad \text{and} \quad \overline{x}_{1,2} = \frac{1}{n-2}\sum_{i=3}^{n} x_i$$

Grubbs [10] proposes this statistic for testing if the two largest or smallest values are outliers. He also provides a table of percentage points for $4 \leq n \leq 149$ and $\alpha = \{0.1\%, 0.5\%, 1\%, 2.5\%, 5\%, 10\%\}$ in [12] , where the observations would be outliers if the statistic is lower than the tabulated percentage for the chosen confidence level.

12 Score Calculations

The different scores are calculated as follows:

1. Normal Scores

$$Z_{score}(i) = \frac{x_i - \overline{x}}{s} \tag{25}$$

where s is the sample standard deviation, $Z_{score} \sim N(0, \frac{n-1}{n})$ according to Pope [19], and $\frac{n-1}{n} \to 1$ for larger n.

2. t-student Scores

$$t_{score}(i) = \frac{Z_{score}(i)\sqrt{n-2}}{\sqrt{n-1-Z^2_{score}(i)}} \tag{26}$$

where $t_{score} \sim t_{n-2}$ according to Grubbs [10].

3. chi-squared Score

$$\mathcal{X}^2_{score}(i) = \frac{(x_i - \overline{x})^2}{s^2} \tag{27}$$

where $\mathcal{X}^2_{score} \sim \mathcal{X}^2(1)$.

4. IQR Score

$$IQR_{score}(i) = \frac{x_i \mathbb{1}_{\{x_i<Q1\}} + x_i \mathbb{1}_{\{x_i>Q3\}} - Q1 \mathbb{1}_{\{x_i<Q1\}} - Q3 \mathbb{1}_{\{x_i>Q3\}}}{Q3 - Q1} \tag{28}$$

where $Q1$ is the 25% quantile and $Q3$ is the 75% quantile.

5. MAD Score

$$MAD_{score}(i) = \frac{x_i - \tilde{x}}{\widetilde{|x - \tilde{x}|}} \tag{29}$$

where $\tilde{x}$ is the median of the data sample. Please notice that this implementation is not corrected by the expected value of the median of absolute deviations about the median as described before.

13 Exponential Smoothing

Exponential smoothing can be used to make a forecast for the time instant $t+1$ based on the last incoming data point p_t, the last forecast value $\hat{p}_t$, and the smoothing factor $0 < \eta < 1$. The model equation reads:

$$\begin{aligned} \hat{p}_1 &= p_1 \\ \hat{p}_{t+1} &= \eta p_t + (1-\eta)\hat{p}_t \end{aligned} \tag{30}$$

Values of η close to one have less of a smoothing effect and give greater weight to recent changes in the data, while values closer to zero have a greater smoothing effect and are less responsive to recent changes. There is no formally correct procedure for choosing η, but it could be possible to determine it by minimizing the sum of $(p_t - \hat{p}_t)^2$.

Kundzewicz et al. [15] implemented a procedure of judging if the newly incoming data point fits the temporal structure that already exists in the data, containing three steps:

- calculate the forecast for the time instant at which the newly observation is taken
- assessment of the variance of the forecast error

- checking if the newly incoming data point does not substantially differ from the forecast value

A data point is considered to be an outlier if:

$$|\frac{p_t - \hat{p}_t}{s_t}| > k \tag{31}$$

where s_t is the standard deviation of the forecast error $e_t = \hat{p}_t - p_t$, and k is the number of standard deviations on the forecast error. By manipulating the values of η and k, we can influence the number of data points falling outside the acceptance limits.

The calculation of s_t could also be performed by exponential smoothing from $p_t - \hat{p}_t$. This approach could be indicated for the case of non-stationary data.

14 Moving Window Filtering Algorithm

Gutierrez and Gregori [20] propose a similar algorithm to Brownlees and Gallo [2] for outlier detection in which a neighborhood of observations, called a filtering window, is necessary to judge the reliability of a single observation. Such a data window can grow and shrink according to data quality and the volatility of the series. The idea behind the algorithm is to assess the validity of a new observation on the basis of its relative distance from a neighborhood of the closest valid past observations.

Let us consider $\{p_i\}_{i=1}^{T}$ be a time series. The procedure to identify outliers is:

$$|p_i - \overline{p}_i(k)| < \alpha s_i(k) + \gamma = \begin{cases} \text{True} & \text{observation } i \text{ is kept,} \\ \text{False} & \text{observation } i \text{ is substituted} \end{cases} \tag{32}$$

where $\overline{p}_i(k)$ and $s_i(k)$ represent the moving average and the moving standard deviation of the previous k values respectively. The parameter α specifies the number of standard deviations acting as a threshold to consider an observation as an outlier, while the role of the parameter γ is to avoid zero variances produced by sequences of k equal values and should be a multiple of the minimum variation allowed for the specific series.

If the observation does not pass the test, it can be removed or substituted by some other value, e.g., the previous observation. For each of the series, the algorithm has to determine the optimal window width k, the parameter γ and the threshold that might be considered to identify an observation as an outlier.

If the observation is removed, the data cleaning algorithm can be run several times for a grid of different values of parameters $\{k, \alpha, \gamma\}$, and the quality of the cleaning can be done by visual inspection of the series graphs.

If the observation is substituted by the previous observation, Gutierrez and Gregori [20] introduce a penalization statistic D in order to assess the real effectiveness of the filtering algorithm:

$$D = \frac{1}{T}\sum_{i=1}^{T} d_i^2 \tag{33}$$

where

$$d_i = \begin{cases} f_i - (p_i + s_i(k)) & \text{if } f_i > p_i + s_i(k) \\ (p_i - s_i(k)) - f_i & \text{if } f_i < p_i - s_i(k) \end{cases}$$

and f_i is the filtered series after substitution. For each of the series the parameter combination $\{k, \alpha, \gamma\}$ that minimizes D needs to be found by, e.g., computing all possible parameter permutations and choosing the one with the lowest D. The best combination does not need to be unique for all series.

In order to reduce the computational intensity of finding the optimal parameter combinations for all series, Gutierrez and Gregori [20] applied hierarchical clustering to identify the most representative series in different groups according to their market behavior.

15 Tests for Non-Normal Distributions

Outlier identification procedures for normal data are relatively easy to use and quite powerful. However, many univariate data sets do not resemble a normal distribution. In these cases, the normal outlier identification techniques can falsely identify extreme observations as outliers. Removing the most extreme observations will tend to distord the data toward symmetry and a closer resemblance to the normal distribution.

Barnett and Lewis [1] (available for purchase) discuss outliers from different non-normal distributions. Here, we only present the results for log-normal and exponential distributions available in Iglewicz and Hoaglin [13].

15.1 Exponential Distribution

The exponential distribution plays a key role for survival and extreme value data because it often approximates such data reasonably well and relatively simple transformations relate the exponential distribution to a number of other distributions. Therefore, techniques for identifying outliers in exponential data can be used after performing an appropriate transformation.

The Cochran procedure for detecting an upper outlier from an exponentially distributed ordered data sample uses the critical region:

$$\frac{x_n}{\sum_{i=1}^{n} x_i} > K \tag{34}$$

where the critical value K for different confidence levels can be obtained in Barnett and Lewis [1].

Kimber [14] adapted the Generalized ESD procedure to develop a test for up to r upper outliers from the exponential distribution. Starting from an ordered sample, the approach first chooses r and then tests the hypothesis that the r largest observations are outliers. If it is rejected, then it checks if the $r-1$ largest number of observations are outliers or not, and so on until the largest observation is checked.

Specifically, the test statistic S_j is defined as:

$$S_j = \frac{x_{n+1-j}}{\sum_{i=1}^{n+1-j} x_i} \quad j = 1, \ldots, r \tag{35}$$

Then, for $j = r, r-1, \ldots, 1$ we ask whether $S_j > s_j$, where s_j is the appropriate critical value from Kimber [14] available for $k \leq 4$ and $n \leq 140$. The largest value of j, say r^*, for which $S_{r^*} > s_{r^*}$ declares the upper r^* as outliers.

For combinations of r, α and n not available in the tables, the following equation is useful for calculating approximate critical values:

$$\binom{n}{j}[(1-s_j)/(1+js_j-s_j)]^{n-j} = \alpha/r \tag{36}$$

and solving for s_j yields

$$s_j = \frac{1-U}{1+(j-1)U} \quad \text{where} \quad U = \left[\frac{\alpha/r}{\binom{n}{j}}\right]^{\frac{1}{n-j}} \tag{37}$$

Kimber [14] also provides a similar procedure to check for the r lower outliers, where

$$S_j^* = \frac{x_{j+1}}{\sum_{i=1}^{j+1} x_i} \quad j = 1, \ldots, r \tag{38}$$

with values for s_j^* provided for $n \leq 200$ and $r \leq 4$.

Barnett and Lewis [1] provide more tests to check for lowest outliers.

15.2 Log-Normal Distribution

If X is a random variable from a log-normal distribution then the logarithms of the observations follow a normal distribution, i.e., $Y = \ln(X)$ has a normal distribution. The inverse also applies, and a more general form of the log-normal distributions also includes a location parameter θ to handle data that are displaced to the right of zero: $X = \theta + e^Y$.

A first approach then is to apply the ln to the original data, which it is assumed to follow a log-normal distribution, and then apply the tests intended for normal distributions.

However, this transformation may not always be satisfactory if the transformed data does not show too much symmetry around the mean or does not resembles a normal distribution. In this case it can be useful to seek a different transformation for the data.

15.3 Transformations to Normality

One objective is to transform the data to improve the symmetry and thus get the transformed data closer to normality.

Let's consider an ordered sample, the depth of the median is defined as $d_1 = (n+1)/2$, the depth of the fourth moment as $d_2 = (\check{d}_1)/2$, the depth of the eights as $d_3 = (\check{d}_2)/2$, and so on. Here, $\check{d}_j$ indicates the greatest integer $\leq d_1$.

For $j \geq 2$, the jth values are the pair of ordered observations with depth d_j from each end, i.e., $x_L = x_{d_j}$ and $x_U = x_{n+1-d_j}$. The difference between $(x_L + x_U)/2$ and $\tilde{x}$ measures skewness.

The family of power transformations $y = \frac{x^p - 1}{p}, p \neq 0$ approaches the function $y = \ln(x)$ as $p \to 0$, and a simpler and equivalent form of this family can be expressed as:

$$y = \begin{cases} x^p & \text{if } p \neq 0 \\ \ln(x) & \text{if } p = 0 \end{cases} \tag{39}$$

A guide for choosing p is to compute p for each pair of letter values from the equation:

$$\frac{x_L + x_U}{2} - \tilde{x} = (1 - p)\Big[\frac{(x_U - \tilde{x})^2 + (\tilde{x} - x_L)^2}{4\tilde{x}}\Big] \tag{40}$$

The exponent to chose will be a round value close to the median of the computed estimates of p. This method works well if the p are close to each other; otherwise it provides only a rough guide to the appropriate transformation.

References

[1] V. Barnett; T. Lewis (1994). Outliers in Statistical Data. Wiley Series in Probability and Mathematical Statistics, 3rd ed.

[2] C.T. Brownlees; G.M. Gallo (2006). Financial Econometric Analysis at Ultra-High Frequency: Data Handling Concerns. Computational Statistics & Data Analysis 51, pp. 2232–2245.

[3] G. Brys; M. Hubert; P.J. Rousseeuw (2005). A Robustification of Independent Component Analysis. Journal of Chemometrics 19(5 − 7), pp. 364–375.

[4] G. Brys; M. Hubert; A. Struyf (2004). A Robust Measure of Skewness. Journal of Computational and Graphical Statistics 13(4), 996–1017.

[5] H.A David; H.O. Hartley; E.S. Pearson (1954). The Distribution of the Ratio, in a Single Normal Sample, of Range to Standard Deviation. Biometrika 41(3), pp. 482–493.

[6] W.J. Dixon (1950). Analysis of Extreme Values. The Annals of Mathematical Statistics 421(4), pp. 488–506.

[7] W.J. Dixon (1951). Ratios Involving Extreme Values. The Annals of Mathematical Statistics 22(1), pp. 68–78.

[8] S.L.R. Ellison; V.J. Barwick; T.J.D. Farrant (2009). Practical Statistics for the Analytical Scientist. A Bench Guide. RSC Publishing, 2nd edition, Cambridge.

[9] M. Frigge; D.C. Hoaglin; B. iglewicz (1989). Some Implementations of the Boxplot. The American Statistician 43(1), pp. 50–54.

[10] F.E. Grubbs (1950). Sample Criteria for Testing Outlying Observations. The Annals of Mathematical Statistics 21(1), pp. 27–58.

[11] F.E. Grubbs (1969). Procedures for Detecting Outlying Observations in Samples. Technometrics 11(1), pp. 1–21.

[12] F.E. Grubbs; G. Beck (1972). Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations. Technometrics 14(4), pp. 847–854.

[13] B. Iglewicz; D.C. Hoaglin (1993). How to Detect and Handle Outliers. ASQC Basic References in Quality Control, vol. 16, Wisconsin.

[14] A.C. Kimber (1982). Tests for Many Outliers in an Exponential Sample. Journal of the Royal Statistical Society, Series C (Applied Statistics), 31(3), pp. 263–271.

[15] Z.W. Kundzewicz et al (1989). Outliers in Groundwater Quality Time Series. Groundwater Management: Quantity and Quality (Proceedings of the Benidorm Symposium, October), IAHS Publication n. 188.

[16] NIST/SEMATECH e-Handbook of Statistical Methods (2010), $http://www.itl.nist.gov/div898/handbook/$.

[17] E.S. Pearson; C. Chandra Sekar (1936). The Efficiency of Statistical tools and a Criterion for the Rejection of Outlying Observations. Biometrika 28, pp. 308–320.

[18] E.S. Pearson; H.O. Hartley (1970). Biometrica Tables for Statisticians, vol. 1, 3rd. ed., Cambridge University Press.

[19] A.J. Pope (1976). The Statistics of Residuals and the Detection of Outliers. NOAA Technical Report NOS 65 NGS1, U.S. Department of Commerce, Rockville.

[20] J.M.Puigvert Gutierrez; J.F. Gregori (2008). Clustering Techniques Applied to Outlier Detection of Financial Market Series Using a Moving Window Filtering Algorithm. ECB Working Paper Series, n. 948, October.

[21] D.B. Rorabacher (1991). Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon Q Parameter and Related Subrange Ratios at the 95 Percent Confidence Level. Analytical Chemistry 83(2), pp. 139–146.

[22] B. Rosner (1983). Percentage Points for a Generalized ESD Many-Outlier Procedure. Technometrics 25(2), pp. 165–172.

[23] S.S. Shapiro (1986). How to Test normality and Other Distributional Assumptions. The ASQC Basic References in Quality Control: Statistical Techniques, vol. 3, Milwaukee, WI.

[24] R.E. Shiffler (1988). Maximum Z Scores and Outliers. The American Statistician 42(1), pp. 79–80.

[25] J.W. Tukey (1977). Exploratory Data Analysis. Addison Wesley.

[26] E. Vanderviere; M. Huber (2004). An Adjusted Boxplot for Skewed Distributions. COMPSTAT'2004 Symposium, Physica-Verlag/Springer.

[27] S. Walfish (2006). A Review of Statistical Outlier Methods. Pharmaceutical Technology, November.