A comparative study on Univariate Outliers Winsorization methods in Data Science Context

Iyad Alkrunz

12/26/2021

Abstract

Keywords Capping; Flooring; Outlier; Quantile-based.

Introduction

Outliers are values in data that differ extremely from a major sample of the data, the presence of outliers can bias the estimates and, as a consequence, significantly reduce the performance and accuracy of a predictable model. The problem of outlier-detection has attracted the attention of many statisticians and data scientists. (References).

The methods of outlier-detection are broadly classified into different classes, namely distribution-based methods, depth-based methods, and density-based methods (Preparata and Shamos, 1988, Dominguesa, et al 2018).

The argument on the handling of outliers is continued between the belief of Tukey (1959) that rejecting outliers indiscriminately is inappropriate, and other various trimming and winsorization techniques. Thus, after detection, outliers are handled in one of three ways: accommodation, omission, or winsorization.

The accommodation is utilized by robust statistical methods in order to resist the effect of outliers on the parameter estimates (Ekezie & Ogu, 2013), which indirectly destroy the conclusions of the study (Hubert et al., 2008, Farcomeni & Ventura, 2010). Trimming of outliers has been well studied, where (Lix and Keselman, 1998; Yusof et al. 2013) have proved its beneficial in terms of robustness, while the type (symmetric or asymmetric) and percentage of trimming have been discussed by (Babu et al. 1999; Wilcox, 2003).

In winsorization, the extreme values are replaced by other appropriate values to reduce the effect of the outliers on the estimation and modeling power (Frey, 2018). The choose of winsorization percentage cut-off point as well as the winsorization statistic are challenging. A poor choice of winsorization percentage will inflated the mean squared errors (MSE) of desired estimators. Thus, it is recommended the choose cut-off point that minimizes the MSE compared to the classical estimator.

In the context of data science, practitioners used different statistics for winsorization, such as mean, median and quantiles *(References)*. To the best of our knowledge, no study has been published dealing with the impact of different winsorization statistics the estimators. This paper investigates the impact of four winsorization statistics viz mean, median, mode and Quantile-based Flooring and Capping technique on the estimates of parameters of three distributions, namely normal, negative binomial and exponential distributions.

Outliers and Winsorization

Sources and Impact of Outliers

Observed variables often contain outliers that differ extremely from a major sample of the data. Some data sets may come from homogeneous groups; others from heterogeneous groups that have different characteristics regarding a specific variable, such as height data not stratified by gender. Outliers can be caused by incorrect measurements, including data entry errors, or by sampling from a different population than the rest of the data (Frost, 2020).

Outliers may cause a negative effect on data analyses such as biasing the estimation, reduce the predictability of constructed model, or it may provide useful information about data when we look into an unusual response to a given study. The data must be evaluated for the presence of outliers before beginning the procedure with the main bulk of data. Thus, outlier detection is an important part of data analysis in the above two cases.

Outliers Detection

There are different ways and methods of identifying outliers, including square root transformation, median absolute deviation, Grubb's test, Ueda's method as explained recently by (Shimizu, 2022). In this paper we are going to use Tukey's method boxplot (Tukey, 1977); due to its popularity and less sensitivity of outliers' existance compare to other tests.

Boxplot is a well-known simple graphical tool to display information about continuous univariate data based on five summaries, namely, median, lower quartile Q_1, upper quartile Q_3, lower extreme, and upper extreme of a data set. It is less sensitive to extreme values of the data than the previous methods using the sample mean and standard variance because it uses quartiles which are resistant to extreme values. The rule of the method is that any value smaller than the lower fence $L_F = Q1 - \nu * IQR$ or larger than the upper fence $U_F = Q3 + \nu * IQR$ is a possible outlier, where ν is the resistance factor and the interquariles range $IQR = Q_3 - Q_1$.

Different values of ν can be considered, but the nominal value is $\nu = 1.5$ (Hoaglin, et al, 1986). Various versions of the boxplot were also proposed (See Abuzaid et al; 2012, Saeger et al; 2016).

The following subsection discusses the treatment of outliers via winsorization.

Winsorization of outliers

There are two common methods for treating outliers in a data set. The first is to remove outliers as a means of trimming the data set. The second method involves replacing the values of outliers with suitable statistic such as mean, median, mode or quantile-based technique as follows:

1. *Replace outliers by mean* : In this technique the outliers are replaced with the arthematic mean $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ of the remaining observations after removing outliers.

2. *Replace outliers by median* : The median value that is the middle value in a ordered remaining observations

$$Q_2 = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})/2 & \text{if } n \text{ is even} \end{cases}$$

is used to replace the detected outliers.

3. *Replace outliers by mode* : The outliers are replaced with the mode value of the remaining observations, which is appears most often in a set of data values.

4.$Quantile-based\ Flooring\ and\ Capping$: in this quantile-based technique, the maximum outliers are replaced with upper fence U_F (capped), and the minimum outliers are replaced with lower fence L_F (floored).

The following section investigates the effect of the four considered winsorization statistics on the performance of parameter estimates for different probability distributions via Monte carlo simulation.

Simulation (Numerical Study)

An R code has been developed and implemented in R Studio environment to generate random data sets from three different probability distributions namely, normal, negative binomial and exponential distribution.

Settings of Data Generation

Data were generated with four different sample sizes, $n = 20$,50,100 and 200, in such a way that $(1-\epsilon)$ of data are generated from the original distribution (P) and the rest ϵ of data are generated from the contamination distribution (Q). Thus, the contaminated data structure can be formulated as $P_\epsilon = (1-\epsilon)P + \epsilon Q$, where ϵ is the contamination level and ϵ =0.05, 0.10 or 0.15.The following three probability distributions are considered:

Normal distribution

For normal random variable, $X \sim N(\mu, \sigma^2)$; data were generated from the standard normal distribution with $\mu = 0$ and $\sigma^=1$. For contamination procedur; the contaminated data were generated from another normal distribution with $\mu = 4$ and $\sigma = 2$.

The maximum likelihood estimator (MLE) of the mean and standard deviation are obtained as the sample mean $\hat{\mu}_{mle} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,and $\hat{\sigma} = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$, respectively.

Negative binomial distribution

Random variable X follows the negative binomial distribution $X \sim NB(k, p)$ with mean $\mu = \frac{k}{p}$ and variance $\sigma^2 = \frac{k(1-p)}{p^2}$ if X is the count of independent Bernoulli trials required to achieve the k^{th} successful trials when the probability of success is a constant p. The probability of $x = n$ trials is $f(X = n) = \binom{n-1}{k-1}p^k(1-p)^{n-k}$. The MLE of p is given by: $\hat{p} = \frac{k}{x+k}$.

For negative binomial random variable; data are generated with parameters $k = 2$ and $p = 0.2$, while the contaminated data are generated from Poisson distribution with $\lambda = 32$, where the probability of k successes is $P(X = k) = \frac{(e^{\lambda}\lambda^k)}{k!}$.

Exponential distribution

The Exponential distribution is the most commonly used model in reliability and life-testing analysis, (i.e $f(x) = \theta e^{-\theta x}$ for $x \geq 0$). The MLE of θ is given by $\hat{\theta} = \frac{1}{\bar{x}}$.
Data were generated with parameter $\theta = 0.5$, and the contaminated data were generated from exponential distribution with $\theta = 0.05$.

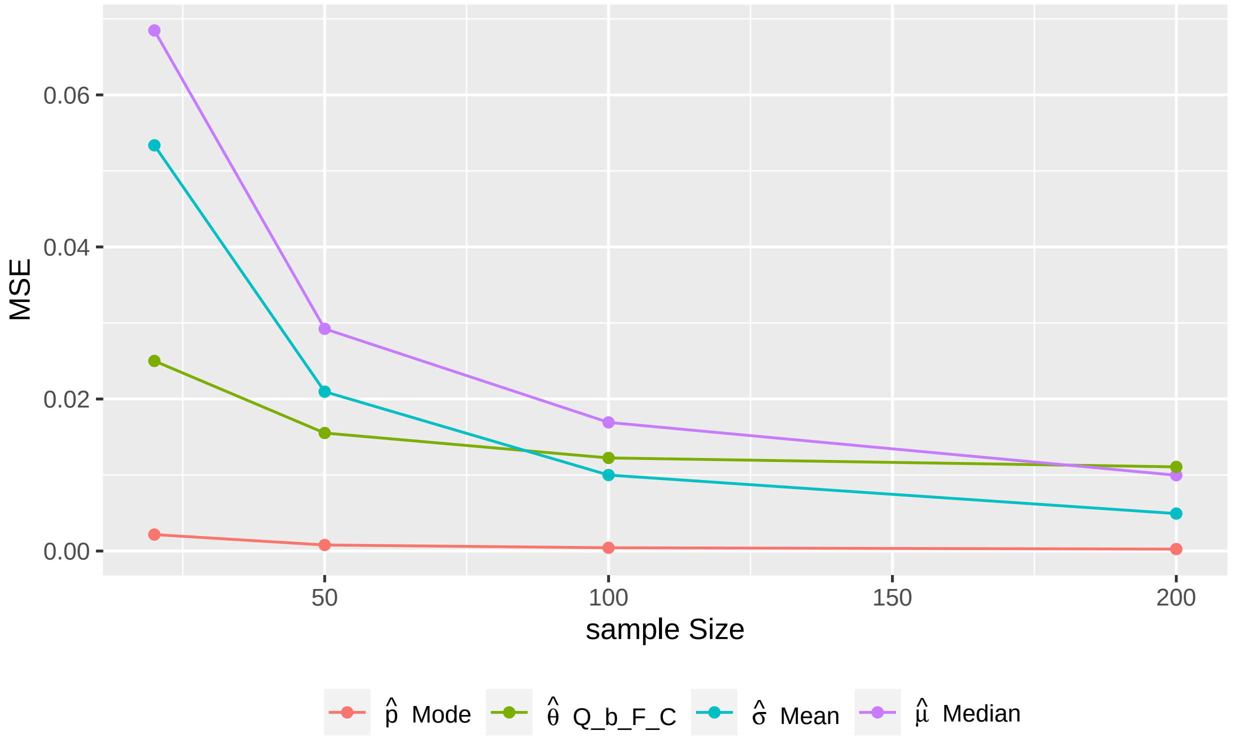
For each combination of the considered probability distributions,sample sizes, contamination levels and winsorization statistics; the generation procedures are repeated 1000 iterations to ensure the convergence.

Performance indicators

The impact of the considered four outliers winsorization statistics on the parameter estimates are measured by three common indicators as follows:

1. *Bias*, is the difference between the estimator's expected value and the true value of the parameter being estimated.
2. *Mean Square Error*(MSE), is a measure of the quality of an estimator. As it is derived from the square of Euclidean distance, it is always a positive value that decreases as the error approaches zero. *Write the formula.*
3. *Goodness of fit tests*, are statistical tests aiming to determine whether a set of observed values match those expected under the applicable distribution. There are different goodness-of-fit tests, in this article the Shipiro-Wilk test is used in the case of normal and exponential distributions, while Kolmogorov-Smirnov test is used in the case of negative binomial distribution.

Results



Regardless the distribution, contamination level or winsorization statistics, the results of simulation studies reveal that, the performance of parameter estimates are improved as the sample size increased, where the MSE has a decreasing function with the sample size, and the bias has a decreasing function of the sample size for n<100 and constant function for n>100, as partially presented in Figure 1.

The performance has relatively an inverse relationship with the contamination level ϵ.

For normal distribution, due to its symmetric nature the mean, median and mode winsorization statistics have almost similar effect on the estimates of the parameters estimates (i.e. μ and σ^2), while they outperform the quantile-based winsorization statistic. For negative binomial case, the mode winsorization statistic outperforms the other winsorization statistics for higher levels of contamination $\epsilon = 0.15$, while the mean winsorization statistic performs better than other winsorization statistics for smaller levels of contamination $\epsilon < 0.15$ For the exponential distribution, the mean winsorization statistic has the best performance followed by median, mode and then the quantile-based method. This behavior may be referred to the MLE estimator of the θ, which is mainly the sample mean.

From the prospective of goodness of fits tests, in the case of normal distribution, mean, median and mode winsorization statistics have consistent performance with respect to the contamination level and sample size, where the proportion of samples fitted by normal distribution close to 1 when the contamination level is $\epsilon = 0.05$. In the case of an exponential distribution, all considered winsorization statistics perform approximately equally and perfectly, where the proportion of samples fitted by exponential distribution close to 1 regardless the sample size or contamination level.
The proportions of fitted sample by negative binomial are less than the other two distributions. The quantile-based winsorization statistic has the worst performance compare to other three considered statistics because it accumulates the winsorizated values at the edges of distribution and malforms the nature of distribution. Thus, the mean winsorization statistic is recommended for most of the cases especially for smaller levels of contaminations.

Normal Distribution

Table 1: Bias (MSE) of the Normal distribution mean estimator for different handling-outliers methods

		Handling-outliers Methods			
n	ϵ	Quantile-based	Mean	Median	Mode
20	5	0.111 (0.06)	0.021 (0.058)	0.02 (0.058)	0.02 (0.058)
50	5	0.092 (0.027)	0.019 (0.021)	0.019 (0.021)	0.019 (0.021)
100	5	0.122 (0.025)	0.029 (0.012)	0.029 (0.012)	0.028 (0.012)
200	5	0.12 (0.02)	0.027 (0.007)	0.027 (0.007)	0.026 (0.007)
20	10	0.24 (0.111)	0.074 (0.068)	0.074 (0.068)	0.074 (0.069)
50	10	0.245 (0.081)	0.068 (0.029)	0.067 (0.029)	0.066 (0.029)
100	10	0.244 (0.07)	0.068 (0.017)	0.066 (0.017)	0.065 (0.017)
200	10	0.245 (0.065)	0.064 (0.01)	0.062 (0.01)	0.061 (0.01)
20	15	0.353 (0.18)	0.113 (0.08)	0.111 (0.08)	0.108 (0.082)
50	15	0.399 (0.182)	0.14 (0.049)	0.136 (0.048)	0.132 (0.048)
100	15	0.378 (0.154)	0.121 (0.028)	0.117 (0.027)	0.113 (0.026)
200	15	0.378 (0.149)	0.119 (0.022)	0.115 (0.021)	0.111 (0.02)

Table 2: Bias (MSE) of the Normal distribution standared deviation estimator for different handling-outliers methods

		Handling-outliers Methods			
n	ϵ	Quantile-based	Mean	Median	Mode
20	5	0.095 (0.047)	0.078 (0.05)	0.077 (0.05)	0.072 (0.049)
50	5	0.094 (0.022)	0.032 (0.017)	0.032 (0.017)	0.029 (0.016)
100	5	0.125 (0.023)	0.026 (0.01)	0.026 (0.01)	0.024 (0.01)
200	5	0.123 (0.019)	0.023 (0.005)	0.023 (0.005)	0.022 (0.005)
20	10	0.226 (0.101)	0.014 (0.053)	0.013 (0.053)	0.007 (0.053)
50	10	0.25 (0.081)	0.005 (0.021)	0.005 (0.021)	0.001 (0.021)
100	10	0.252 (0.073)	0.006 (0.01)	0.006 (0.01)	0.009 (0.01)
200	10	0.255 (0.07)	0.005 (0.005)	0.005 (0.005)	0.007 (0.005)
20	15	0.35 (0.183)	0.004 (0.064)	0.005 (0.064)	0.015 (0.065)
50	15	0.401 (0.188)	0.062 (0.036)	0.063 (0.036)	0.069 (0.036)
100	15	0.387 (0.162)	0.048 (0.016)	0.049 (0.016)	0.053 (0.016)
200	15	0.392 (0.159)	0.055 (0.01)	0.055 (0.01)	0.059 (0.01)

Table 3: The proportion of samples are fitted by normal distribution at 0.05 level of significance after handling outliers.

		Handling-outliers Methods			
n	ϵ	Quantile-based	Mean	Median	Mode
20	5	0.961	0.970	0.949	0.924
50	5	0.926	0.970	0.965	0.958
100	5	0.614	0.968	0.960	0.934
200	5	0.137	0.962	0.953	0.916
20	10	0.873	0.951	0.939	0.910
50	10	0.474	0.938	0.918	0.879
100	10	0.060	0.890	0.873	0.805
200	10	0.000	0.777	0.752	0.664
20	15	0.743	0.937	0.908	0.868
50	15	0.108	0.785	0.737	0.674
100	15	0.006	0.666	0.617	0.540
200	15	0.000	0.251	0.217	0.146

Negative Binomial Distribution

in Negative Binomial distribution we focused on probability of success, we calculated bias and mean squared error for estimator after applying the four different methods of handling outliers, with different sample sizes and different contamination levels, also we applied Goodness of fit test namely Kolmogorov-Smirnov Test to determines if sample follows a Negative Binomial distribution

Table 4: Bias (MSE) of the Negative Binomial Distribution probability of success estimator for different handling-outliers methods

		Handling-outliers Methods			
n	ϵ	Quantile-based	Mean	Median	Mode
20	5	0.013 (0.001)	0.017 (0.002)	0.018 (0.002)	0.019 (0.002)
50	5	0.014 (0)	0.011 (0.001)	0.012 (0.001)	0.014 (0.001)
100	5	0.017 (0)	0.01 (0)	0.011 (0)	0.014 (0.001)
200	5	0.018 (0)	0.008 (0)	0.01 (0)	0.013 (0)
20	10	0.031 (0.001)	0.005 (0.002)	0.007 (0.002)	0.009 (0.002)
50	10	0.034 (0.001)	0.001 (0.001)	0.003 (0.001)	0.006 (0.001)
100	10	0.034 (0.001)	0.002 (0)	0 (0)	0.004 (0)
200	10	0.034 (0.001)	0.001 (0)	0.001 (0)	0.006 (0)
20	15	0.045 (0.002)	0.006 (0.002)	0.004 (0.002)	0.001 (0.002)
50	15	0.049 (0.002)	0.019 (0.001)	0.017 (0.001)	0.015 (0.001)
100	15	0.047 (0.002)	0.015 (0.001)	0.013 (0.001)	0.009 (0.001)
200	15	0.046 (0.002)	0.016 (0)	0.014 (0)	0.01 (0)

Table 5: The proportion of samples are fitted by Negative binomial distribution at 0.05 level of significance after handling outliers.

		Handling-outliers Methods			
n	ϵ	Quantile-based	Mean	Median	Mode
20	5	0.748	0.912	0.878	0.848
50	5	0.442	0.682	0.606	0.538
100	5	0.294	0.450	0.418	0.356
200	5	0.152	0.254	0.200	0.198
20	10	0.582	0.846	0.792	0.782
50	10	0.256	0.542	0.488	0.380
100	10	0.090	0.352	0.314	0.234
200	10	0.008	0.258	0.192	0.110
20	15	0.514	0.770	0.720	0.700
50	15	0.128	0.428	0.346	0.296
100	15	0.034	0.218	0.168	0.150
200	15	0.000	0.146	0.116	0.064

Exponential distribution

Exponential distribution has only one parameter λ, we calculated bias and mean squared error for estimator after applying the four different methods of handling outliers, with different sample sizes and different contamination levels, also we applied Goodness of fit test namely Chi-Square Test to determines if sample follows a Exponential distribution

Table 6: Bias (MSE) of the exponential distribution Rate estimator for different handling-outliers methods

			Handling-outliers Methods			
n	ϵ	Before	Quantile-based	Mean	Median	Mode
20	5	0.359 (0.13)	0.069 (0.017)	0.077 (0.033)	0.092 (0.037)	0.115 (0.045)
50	5	0.309 (0.096)	0.039 (0.007)	0.08 (0.018)	0.093 (0.021)	0.116 (0.028)
100	5	0.335 (0.113)	0.045 (0.004)	0.07 (0.01)	0.082 (0.012)	0.107 (0.018)
200	5	0.329 (0.109)	0.043 (0.003)	0.066 (0.007)	0.078 (0.009)	0.104 (0.014)
20	10	0.382 (0.147)	0.13 (0.025)	0.055 (0.024)	0.076 (0.029)	0.106 (0.038)
50	10	0.376 (0.142)	0.108 (0.016)	0.052 (0.012)	0.07 (0.015)	0.103 (0.022)
100	10	0.375 (0.141)	0.102 (0.012)	0.051 (0.007)	0.069 (0.009)	0.103 (0.016)
200	10	0.374 (0.14)	0.101 (0.011)	0.047 (0.004)	0.063 (0.006)	0.098 (0.012)
20	15	0.395 (0.157)	0.179 (0.038)	0.038 (0.018)	0.064 (0.022)	0.1 (0.032)
50	15	0.38 (0.145)	0.151 (0.026)	0.042 (0.01)	0.065 (0.013)	0.103 (0.022)
100	15	0.393 (0.155)	0.159 (0.026)	0.03 (0.004)	0.053 (0.007)	0.096 (0.014)
200	15	0.393 (0.154)	0.156 (0.025)	0.029 (0.003)	0.051 (0.005)	0.095 (0.012)

Table 7: The proportion of samples are fitted by Exponential distribution at 0.05 level of significance after handling outliers.

			Handling-outliers Methods			
n	ϵ	Before	Quantile-based	Mean	Median	Mode
20	5	0.000	0.999	1.000	0.996	0.984
50	5	0.000	1.000	1.000	1.000	0.997
100	5	0.000	1.000	1.000	1.000	1.000
200	5	0.000	1.000	1.000	1.000	1.000
20	10	0.000	0.997	0.998	0.986	0.964
50	10	0.000	0.998	0.999	0.996	0.984
100	10	0.000	0.998	1.000	1.000	0.998
200	10	0.000	1.000	1.000	1.000	0.997
20	15	0.002	0.994	0.994	0.968	0.920
50	15	0.000	0.992	1.000	0.990	0.951
100	15	0.000	0.988	0.999	0.995	0.945
200	15	0.000	0.982	1.000	0.999	0.944

REFERENCES

Abuzaid, A. H., Mohamed, I. B. and Hussin, A. G. (2012). Boxplot for Circular Variables. Computational Statistics. 27 (3), 381-392.

Saeger, T. Kleven, B. Otero, I., Wallace, M. and Ziglar, R.(2016). Outlier Labeling Method for Univariate Data for Module Test and Die Sort. IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING, 29 (4), 330-335.

Nyitrai,T. and Miklos, M. (2019) The effects of handling outliers on the performance of bankruptcy prediction models. Socio-Economic Planning Sciences, 67, 34-42.

Babu GJ, Padmanabhan AR, and Puri ML (1999). Robust One-way ANOVA under Possibly Non Regular Conditions. Biometrical Journal, 41: 321-339.

Dominguesa R, Filipponea M, Michiardia P, Zouaouib J. A Comparative Evaluation of Outlier Detection Algorithms: Experiments and Analyses, Pattern Recogn. 2018; 74: 406-421.

Ekezie DD and Ogu AI (2013). Statistical Analysis/Methods of Detecting Outliers in Univariate Data in A Regression Analysis Model. International Journal of Education and Research, 1(5): 1-24.

Hubert M, Rousseeuw PJ, and Van Aelst S (2008). High-breakdown Robust Multivariate Methods. Statistical Science, 23(1): 92-119.

Lix LM and Keselman HJ (1998). To Trim or Not to Trim: Tests of Location Equality under Heteroscedasticity and Non-normality. Educational and Psychological Measurement, 115: 335-363.

Preparata F, Shamos M.Computational Geometry: an Introduction, Springer-Verlag, Berlin;1988.

Wilcox RR (2003). Applying Contemporary Statistical Techniques. Academic Press: San Diego, CA.

Yusof ZM, Othman AR, and Syed Yahaya SS (2013). Robustness of Trimmed F Statistics when Handling Nonnormal Data. Malaysian Journal of Science, 32(1): 73-77.

Frey, B. (2018). The SAGE encyclopedia of educational research, measurement, and evaluation (Vols. 1-4). Thousand Oaks,, CA: SAGE Publications, Inc. doi: 10.4135/9781506326139

Tukey, J. W. (1959). A survey of sampling from contaminated distributions. Princeton, New Jersey: Princeton University.

Seo, S.(2006) A review and comparison of methods for detecting outliers in univariate data sets. Diss. University of Pittsburgh.

Kwak, Sang Kyu, and Jong Hae Kim. “Statistical data preparation: management of missing values and outliers.” Korean journal of anesthesiology 70.4 (2017): 407.

Frost J (2020). Hypothesis testing: An intuitive guide for making data drives decisions. Statistics by Jim Publishing State College, Pennysalvia, U.S.A.

Shimizu Y (2022) Multiple Desirable Methods in Outlier Detection of Univariate Data With R Source Codes. Front. Psychol. 12:819854.

Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading.

Hoaglin DC, Iglewicz B, Tukey JW (1986) Performance of some resistant rules for outlier labeling. J Am Stat Assoc 81(396):991-999

https://www.redalyc.org/pdf/2990/299023509004.pdf