doi.org/10.26398/IJAS.0036-04

A COMPARATIVE STUDY ON UNIVARIATE OUTLIER WINSORIZATION METHODS IN DATA SCIENCE CONTEXT

Ali Abuzaid[1]

Department of Mathematics, Al Azhar University - Gaza, Gaza, Palestine.

Iyad Alkrunz

Department of Information Technology, Al Azhar University - Gaza, Gaza, Palestine.

Abstract *Handling outliers is an important step in data analysis, and it can be approached through three different ways, namely; accommodation, omission, or winsorization. This article investigates the impact of four winsorization statistics (mean, median, mode, and quantiles) on parameter estimation through an extensive simulation study. Three probability distributions (normal, negative binomial, and exponential) are considered, each with varying degrees of contamination. The simulation results suggest that winsorization is effective for small contamination levels and large sample sizes. Furthermore, it is recommended to winsorize outliers in symmetric distributions using any of the location parameters. However, for asymmetric distributions, the median should be employed. To illustrate these findings, a real dataset on internet usage session durations for 4,500 users, comprising over 2 million records, are fitted to the exponential distribution. The identified outliers were winsorized using the aforementioned statistics.*

Keywords: *Capping; flooring; outlier; quantile-based.*

1. Introduction

Outliers refer to data values that significantly deviate from the majority of the data. The presence of outliers can have a detrimental impact on the effectiveness and accuracy of a predictive model, as they have the potential to skew estimations. Outliers can arise due to various factors, such as incorrect measurements, data entry errors, or sampling from a different population (Frost, 2020). Consequently, the issue of outlier-detection has garnered considerable attention from statisticians and data scientists.

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).

[1] Email: a.abuzaid@alazhar.edu.ps

The argument on the handling of outliers is continued between the belief of Tukey (1960) that rejecting outliers indiscriminately is inappropriate, and other various trimming and winsorization techniques. Thus, after detection, outliers can be handled in one of three ways: accommodation, omission, or winsorization.

Accommodation is employed by robust statistical methods to mitigate the impact of outliers on parameter estimates (Ekezie and Ogu, 2013). Outliers have the potential to undermine the conclusions of a study (Hubert et al., 2008; Farcomeni and Ventura, 2010), and thus, accommodation techniques are utilized to indirectly counteract their influence. The trimming of outliers has been extensively stud-ied, and researchers such as Lix and Keselman (1998) and Yusof et al. (2013) have proved its benefits in terms of improving robustness. Additionally, the topic of trimming, including discussions on the type (symmetric or asymmetric) and percentage of trimming, has been addressed by Babu et al. (1999) and Wilcox (2003).

In winsorization, extreme values are substituted with suitable values to mitigate the impact of outliers on estimators and modeling power (Frey, 2018). These substitute values can be any of the central tendency measures as outlined in Section 2. However, determining the appropriate winsorization percentage cut-off point and the winsorization statistic can pose challenges.

A poor choice of winsorization percentage will inflate the mean squared errors (MSE) of desired estimators. Thus, it is recommended to choose the cut-off point that minimizes the MSE compared to the classical estimator. Winsorization is recommended to avoid the loss of power (Leys, et al, 2019). Moreover, Liao et al (2017) highlighted the effectiveness of winsorization in controlling Type I error inflation and outlier impact on power based on a simulation study.

In the context of data science, practitioners used different statistics for winsorization, such as mean, median and quantiles. To the best of our knowledge, no published study has specifically examined the impact of different winsorization statistics on estimators. This article 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 distribution.

2. Outliers and Winsorization

2.1. Outliers Detection

Various methods exist for identifying outliers, such as square root transformation, median absolute deviation, Grubb’s test, and Ueda’s method, as recently discussed by Shimizu (2022). However, in this article, we use Tukey's method boxplot (1977) due to its popularity and less sensitivity of outliers' existence compared 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. Any value smaller than the lower fence $L_F = Q_1 - \nu * IQR$ or larger than the upper fence $U_F = Q_3 + \nu * IQR$ is an outlier candidate, where ν is the resistance factor and $IQR = Q_3 - Q_1$ is the interquartile range. 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.

2.2. Winsorization of outliers

The winsorization method involves replacing outlier values with a suitable statistic such as mean, median, mode or quantile-based technique as follows:

1. $Replacing\ outliers\ by\ mean$: In this technique, outliers are replaced with the arithmetic mean of the remaining observations after removing outliers.

2. $Replacing\ outliers\ by\ median$: The median value, which is the middle value of an ordered remaining observations, is used to replace the detected outliers.

3. $Replacing\ outliers\ by\ mode$: Outliers are replaced with the mode value of the remaining observations.

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

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

3. Simulation

An *R* code has been developed to generate random datasets from three different probability distributions, namely, normal, negative binomial and exponential distribution.

3.1. Settings of Data Generation

Data were generated with four different sample sizes, n = 20, 50, 100 and 200, in such a way that $(1 - \varepsilon)$ of data are generated from the original distribution (P) and the rest ε of data are generated from the contamination distribution (C). Thus, the contaminated data structure can be formulated as $P_\varepsilon = (1-\varepsilon)P+\varepsilon C$, where ε is the contamination level and ε = 0.05, 0.10, 0.15 and 0.20. The following three probability distributions are considered:

3.1.1. Normal distribution

Let X be a random variable having the normal distribution, with mean $-\infty < \mu < \infty$ and standard deviation $\sigma > 0$, $X \sim N(\mu, \sigma^2)$. The datasets were generated from the standard normal distribution with $\mu = 0$ and $\sigma = 1$. For contamination procedure, 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 $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $\hat{\sigma} = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$, respectively. Moreover, the least squares estimation method is equivalent to the MLE, where both are sensitive to the presence of outliers.

3.1.2. Negative binomial distribution

Let X be a random variable having the negative binomial distribution, $X \sim NB(k, p)$ with mean, $\mu = \frac{k}{p}$ and variance $\sigma^2 = \frac{k(1-p)}{p^2}$, where X is the count of independent Bernoulli trials are required to achieve the k^{th} successful trials when the probability of success is a constant p, and $p \in [0,1]$. The probability of $f(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$, for $x = k, k+1, k+2, \ldots$ and $k = 0, 1, 2, \ldots$ The MLE of p is given by $\hat{p} = \frac{k}{x+k}$.

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

3.1.3. Exponential distribution

The exponential distribution is the most commonly used model in reliability and life-testing analysis. The probability density function of a random variable X, having the exponential distribution is given by $f(x) = \theta e^{-\theta x}$ for $x \geq 0$ and $\theta > 0$.

The *MLE* of θ is given by $\hat{\theta} = \frac{1}{\bar{x}}$, where $\bar{x}$ is the sample mean.
Data were generated from the exponential distribution with parameter $\theta = 0.5$, and the contaminated data were generated from exponential distribution with $\theta = 0.05$.

For each combination of probability distributions, sample sizes, contamination levels and winsorization statistics, the generation procedure is repeated 1000 iterations to ensure the convergence.

3.2. Results

The impact of the four outliers winsorization statistics on the parameter estimators is measured by three common indicators as follows:

1. *Bias*, it is the difference between the estimator's expected value and the true value of the parameter being estimated.

2. $Mean\ Square\ Error$, $MSE = \frac{1}{1000}\sum_{i=1}^{1000}(\beta - \hat{\beta}_i)^2$, where β and $\hat{\beta}_i$ are the true and estimated values of the considered parameters.

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 Shapiro-Wilk test is used in the case of normal distribution and exponential distri-bution, while the Kolmogorov-Smirnov test is used in the case of negative binomial distribution.

The simulation results are summarized in Tables (1-5). Regardless of the distribution, contamination level, or winsorization statistics employed, the simulation study reveals that the performance of parameter estimators improves with larger sample sizes. Specifically, the mean squared error (MSE) and bias exhibit an inverse relationship with the sample size ($n < 100$), while they stabilize as a constant function for $n \geq 100$. This relationship is partially illustrated in Figure 1.

The performance has a relatively inverse relationship with the contamination level (ε).

For the normal distribution, due to its symmetric nature, the mean, median and mode winsorization statistics have an almost similar effect on the estimators of the parameters (i.e., μ and σ^2), while they outperform the quantile-based winsorization statistic as given in Tables (1-2).

For the negative binomial distribution, the mode winsorization statistic slightly outperforms the other winsorization statistics for higher levels of contamination

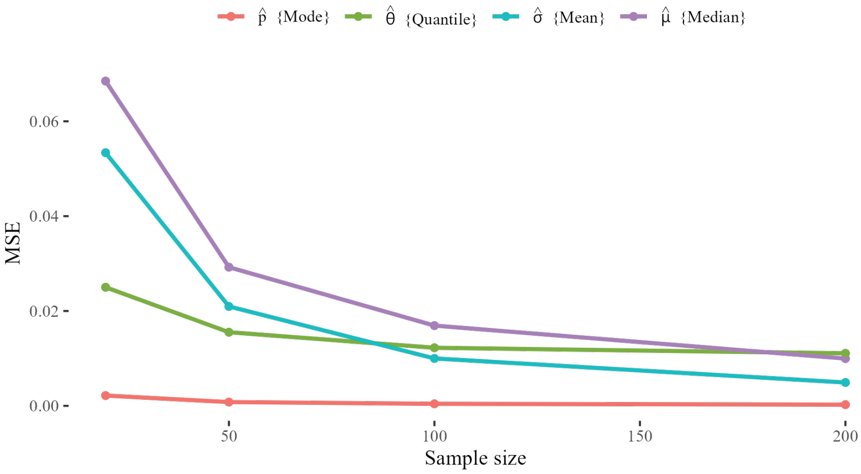

Figure 1: MSE of different parameters' estimators after using winsorization methods for different sample sizes

$(\varepsilon \geq 0.15)$, while the mean winsorization statistic performs better than other winsorization statistics for smaller levels of contamination $(\varepsilon < 0.15)$ as presented in Table 3.

For the exponential distribution (Table 4), the mean winsorization statistic has the best performance, followed by the median, mode and then the quantile-based method. This behavior may be referred to the properties of the *MLE* estimator of the parameter (θ), which is mainly the sample mean.

Table 5 presents the proportion of fitted samples by the associated distributions at 0.05 level of significance before winsorization $(\varepsilon = 0)$, where the proportions are close to 0.95 for the considered sample sizes and probability distributions. The proportions of fitted samples have an inverse relationship with the contamination level. The quantile-based winsorization statistic has the worst performance compared to the other three considered statistics because it accumulates the winsorizated values at the edges of the distribution and malforms the nature of the distribution. Thus, the mean winsorization statistic is recommended for most of the cases, especially for smaller levels of contamination $(\varepsilon \leq 0.1)$.

For normal distribution, mean, median and mode winsorization statistics have consistent performance with respect to the contamination level and sample size,

Table 1: Bias (MSE) of the normal distribution's mean estimator for different winsorization methods

		Winsorization methods			
n	ε	Quantile-based	Mean	Median	Mode
20	0	0.005 (0.053)	0.005 (0.059)	0.005 (0.059)	0.005 (0.059)
50	0	0 (0.021)	0.001 (0.023)	0.001 (0.023)	0.002 (0.023)
100	0	0.003 (0.01)	0.004 (0.01)	0.004 (0.01)	0.004 (0.01)
200	0	0.001 (0.005)	0.001 (0.005)	0.001 (0.005)	0.001 (0.005)
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)
20	20	0.489 (0.293)	0.209 (0.118)	0.204 (0.115)	0.204 (0.115)
50	20	0.497 (0.271)	0.189 (0.067)	0.183 (0.064)	0.179 (0.062)
100	20	0.509 (0.271)	0.193 (0.054)	0.186 (0.051)	0.179 (0.049)
200	20	0.515 (0.271)	0.197 (0.047)	0.19 (0.044)	0.184 (0.042)

where the proportions of the fitted samples by normal distribution are close to 1 when the contamination level is $(\varepsilon = 0.05)$. In the case of an exponential distribution, all considered winsorization statistics perform approximately the same, where the proportions of fitted samples by exponential distribution are close to 1 regardless the sample size or contamination level.

The proportions of fitted samples by negative binomial distribution are less than the other two distributions.

4. Application

A dataset on internet usage was obtained from the Ministry of Telecom and Information Technology in Palestine. The dataset comprises more than 2 mil-

Table 2: Bias (MSE) of the normal distribution's standard deviation estimator for different winsorization methods

		Winsorization methods			
n	*ε*	Quantile-based	Mean	Median	Mode
20	0	0.037 (0.027)	0.074 (0.042)	0.074 (0.042)	0.073 (0.042)
50	0	0.02 (0.011)	0.054 (0.017)	0.054 (0.017)	0.053 (0.017)
100	0	0.011 (0.005)	0.039 (0.009)	0.039 (0.009)	0.039 (0.008)
200	0	0.009 (0.003)	0.037 (0.005)	0.037 (0.005)	0.037 (0.005)
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)
20	20	0.487 (0.307)	0.131 (0.109)	0.132 (0.11)	0.142 (0.111)
50	20	0.514 (0.295)	0.125 (0.054)	0.126 (0.054)	0.132 (0.056)
100	20	0.526 (0.291)	0.129 (0.035)	0.13 (0.035)	0.135 (0.037)
200	20	0.532 (0.29)	0.137 (0.028)	0.137 (0.028)	0.141 (0.029)

Table 3: Bias (MSE) of the negative binomial distribution probability of success estimator for different winsorization methods

		Winsorization methods			
n	ε	Quantile-based	Mean	Median	Mode
20	0	0.006 (0.001)	0.019 (0.002)	0.02 (0.002)	0.021 (0.002)
50	0	0.004 (0.000)	0.016 (0.001)	0.017 (0.001)	0.018 (0.001)
100	0	0.003 (0.000)	0.015 (0.000)	0.016 (0.001)	0.017 (0.001)
200	0	0.002 (0.000)	0.013 (0.000)	0.014 (0.000)	0.015 (0.000)
20	5	0.013 (0.001)	0.017 (0.002)	0.018 (0.002)	0.019 (0.002)
50	5	0.014 (0.000)	0.011 (0.001)	0.012 (0.001)	0.014 (0.001)
100	5	0.017 (0.000)	0.01 (0.000)	0.011 (0.000)	0.014 (0.001)
200	5	0.018 (0.000)	0.008 (0.000)	0.01 (0.000)	0.013 (0.000)
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 (0.000)	0.000 (0.000)	0.004 (0.000)
200	10	0.034 (0.001)	0.001 (0.000)	0.001 (0.000)	0.006 (0.000)
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 (0.000)	0.014 (0.000)	0.01 (0.000)
20	20	0.056 (0.003)	0.03 (0.002)	0.029 (0.002)	0.027 (0.002)
50	20	0.056 (0.003)	0.034 (0.002)	0.032 (0.002)	0.03 (0.002)
100	20	0.056 (0.003)	0.037 (0.002)	0.035 (0.002)	0.032 (0.002)
200	20	0.056 (0.003)	0.04 (0.002)	0.038 (0.002)	0.036 (0.002)

Table 4: Bias (MSE) of the exponential distribution's rate estimator for different winsorization methods

		Winsorization Methods			
n	ε	Quantile-based	Mean	Median	Mode
20	0	0.004 (0.016)	0.116 (0.049)	0.127 (0.054)	0.147 (0.064)
50	0	0.013 (0.006)	0.101 (0.022)	0.11 (0.025)	0.127 (0.031)
100	0	0.018 (0.003)	0.094 (0.015)	0.102 (0.017)	0.118 (0.021)
200	0	0.023 (0.002)	0.092 (0.012)	0.1 (0.013)	0.116 (0.017)
20	5	0.069 (0.017)	0.077 (0.033)	0.092 (0.037)	0.115 (0.045)
50	5	0.039 (0.007)	0.08 (0.018)	0.093 (0.021)	0.116 (0.028)
100	5	0.045 (0.004)	0.07 (0.01)	0.082 (0.012)	0.107 (0.018)
200	5	0.043 (0.003)	0.066 (0.007)	0.078 (0.009)	0.104 (0.014)
20	10	0.13 (0.025)	0.055 (0.024)	0.076 (0.029)	0.106 (0.038)
50	10	0.108 (0.016)	0.052 (0.012)	0.07 (0.015)	0.103 (0.022)
100	10	0.102 (0.012)	0.051 (0.007)	0.069 (0.009)	0.103 (0.016)
200	10	0.101 (0.011)	0.047 (0.004)	0.063 (0.006)	0.098 (0.012)
20	15	0.179 (0.038)	0.038 (0.018)	0.064 (0.022)	0.1 (0.032)
50	15	0.151 (0.026)	0.042 (0.01)	0.065 (0.013)	0.103 (0.022)
100	15	0.159 (0.026)	0.03 (0.004)	0.053 (0.007)	0.096 (0.014)
200	15	0.156 (0.025)	0.029 (0.003)	0.051 (0.005)	0.095 (0.012)
20	20	0.227 (0.057)	0.035 (0.019)	0.066 (0.024)	0.11 (0.036)
50	20	0.213 (0.048)	0.029 (0.008)	0.058 (0.011)	0.11 (0.022)
100	20	0.211 (0.045)	0.018 (0.004)	0.046 (0.006)	0.097 (0.014)
200	20	0.209 (0.044)	0.016 (0.002)	0.044 (0.004)	0.097 (0.012)

Table 5: The proportion of fitted samples by associated distributions at 0.05 level of significance.

Distribution		Normal distribution				Exponential distribution				Negative binomial distribution			
n	ε	Qun.	Mean	Med	Mode	Qun.	Mean	Med.	Mode	Qun.	Mean	Med.	Mode
20	0	(0.970)				(0.964)				(0.922)			
50	0	(0.964)				(0.935)				(0.930)			
100	0	(0.944)				(0.939)				(0.924)			
200	0	(0.951)				(0.941)				(0.938)			
20	5	0.961	0.970	0.949	0.924	0.999	1.000	0.996	0.984	0.748	0.912	0.878	0.848
50	5	0.926	0.970	0.965	0.958	1.000	1.000	1.000	0.997	0.442	0.682	0.606	0.538
100	5	0.614	0.968	0.960	0.934	1.000	1.000	1.000	1.000	0.294	0.450	0.418	0.356
200	5	0.137	0.962	0.953	0.916	1.000	1.000	1.000	1.000	0.152	0.254	0.200	0.198
20	10	0.873	0.951	0.939	0.910	0.997	0.998	0.986	0.964	0.582	0.846	0.792	0.782
50	10	0.474	0.938	0.918	0.879	0.998	0.999	0.996	0.984	0.256	0.542	0.488	0.380
100	10	0.060	0.890	0.873	0.805	0.998	1.000	1.000	0.998	0.090	0.352	0.314	0.234
200	10	0.000	0.777	0.752	0.664	1.000	1.000	1.000	0.997	0.008	0.258	0.192	0.110
20	15	0.743	0.937	0.908	0.868	0.994	0.994	0.968	0.920	0.514	0.770	0.720	0.700
50	15	0.108	0.785	0.737	0.674	0.992	1.000	0.990	0.951	0.128	0.428	0.346	0.296
100	15	0.006	0.666	0.617	0.540	0.988	0.999	0.995	0.945	0.034	0.218	0.168	0.150
200	15	0.000	0.251	0.217	0.146	0.982	1.000	0.999	0.944	0.000	0.146	0.116	0.064
20	20	0.551	0.839	0.802	0.740	0.956	0.989	0.951	0.874	0.414	0.646	0.580	0.592
50	20	0.043	0.631	0.576	0.518	0.931	0.998	0.988	0.866	0.134	0.316	0.250	0.196
100	20	0.000	0.256	0.226	0.173	0.914	1.000	0.994	0.831	0.014	0.134	0.112	0.072
200	20	0.000	0.021	0.017	0.010	0.817	1.000	0.998	0.738	0.002	0.032	0.026	0.008

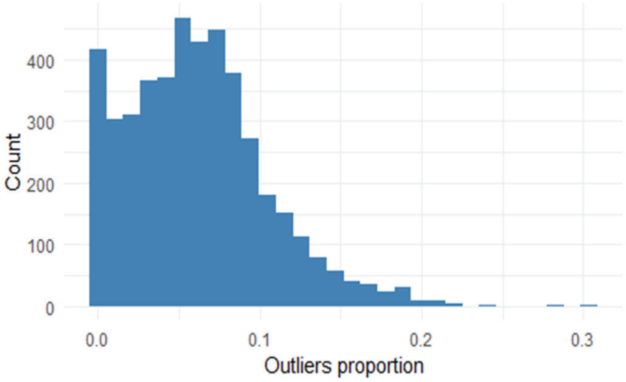

Figure 2: Histogram of detected outliers proportion

lion session records for 4,500 randomly selected users from an internet service provider company in Palestine. Each session in the dataset includes various features such as start-time, end-time, traffic, and duration.

In this example, we are interested only in sessions' durations, which are commonly hypothesized to be exponentially distributed (see Akmeroth and Ammaram, 1996, Sripanidkulchai, et al, 2004, Chetlapalli, et al, 2020). Consonance with that, we assume that sessions' duration are exponentially distributed; therefore, the sessions rows are aggregated for each user. A total of 1,416 (31.467%) of user sessions' duration have been fitted by exponential distribution at 0.05 level of significance according to Shapiro-Wilk goodness-of-fit test.

The outliers of sessions' duration for each user have been detected. Figure 2 presents the proportions of detected outliers for each user, it ranges between 0% and 30%, with mean of 6% and it is an obvious positively skewed distribution.

Three winsorization methods are applied to users' sessions duration data, which are identified as outliers. The summary of fitted users before and after winsorization is presented in Table 6. The results show that the proportion of fitted users data after winsorization is increased significantly, where the mean has the highest proportion, followed by the median and then the quantile-based method which are consistent with the findings of the simulation study. The Chi-square test of independence shows that there are significant associations between the status of users' sessions duration data (i.e fitted by exponential distribution) before and after winsorization at 0.05 level of significance. These associations reveal that an insignificant number of the exponentially fitted users data before winsorization

Table 6: Summary of fitted users data before and after winsorization by exponential distribution

		After Outliers Winsorization		
Statistics	Before	Mean	Median	Quantile-based
Proportion of fit	0.31	0.67	0.61	0.56
Chi-square test	-	1011.42	1311.84	1632.54
p-value	-	0.00	0.00	0.00

has been alternated to be not fitted by exponential after winsorization has been conducted.

5. Conclusions

The winsorization techniques to handle outliers in univariate data have been evaluated via a simulation study. The findings revealed that the nature of the data, including its distribution shape, sample size and contamination level, are the key factors. Thus, it is recommended to use winsorization techniques for large samples $(n \geq 100)$ with a small level of contamination $(\varepsilon \leq 0.1)$. In the case of symmetric distributions, any of the central tendency measures can be used, while for the asymmetric distributions, the use of the median is recommended. This article has focused on three commonly used probability distributions: normal, negative binomial, and exponential. However, further studies could explore other univariate and multivariate distributions to gain a more comprehensive understanding. Additionally, robust statistics remain a viable alternative to winsorization, and it would be valuable to compare their performance in outlier handling techniques.

References

Abuzaid, AH., Mohamed, IB. and Hussin, A.G. (2012). Boxplot for circular variables. *Computational Statistic*s. 27 (3), 381-392.

Almeroth, KC., and Ammarm, MH. (1996). Collecting and modeling the join/leave behavior of multicast group members in the MBone. In *Proceedings of International Symposium on High Performance Distributed Computing* (HPDC).

Babu, G.J., Padmanabhan, A.R. and Puri ML (1999). Robust one-way ANOVA under possibly non regular conditions. *Biometrical Journal*, 41: 321-339.

Chetlapalli, V., Iyer, K.S.S. and Agrawal, H. (2020) Modelling time-dependent aggregate traffic in 5G networks. *Telecommun Syst* 73, 557-575. https://doi.org/10.1007/s11235-019-00629-w

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

Ekezie, D.D., and Ogu, A.I. (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.

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

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

Hoaglin, D.C., Iglewicz B. and Tukey, J.W. (1986) Performance of some resistant rules for outlier labeling. *J Am Stat Assoc* 81(396):991-999

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

Kwak, S.K. and Kim, J.H. (2017) Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology* 70(4): 407-411.

Leys, C., Delacre, M., Mora, Y.L., Lakens, D. and Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on preregistration. *International Review of Social Psychology*, 32(1): 5, 1-10.

Liao, H, Yanju, Li and Brooks, GP (2017). Outlier impact and accommodation on power. *Journal of Modern Applied Statistical Methods*, 16(1): 261-278.

Lix, L.M. and Keselman, H.J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and non-normality. *Educational and Psychological Measurement*, 115: 335-363.

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.

Preparata, F., Shamos, M. (1988) *Computational Geometry: An Introduction*, Springer-Verlag, Berlin.

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.

Shimizu, Y. (2022) Multiple desirable methods in outlier detection of univariate data with R source codes. *Front. Psychol*. 12:819854.

Sripanidkulchai, K., Maggs, B. and Zhang, H. (2004). An analysis of live streaming workloads on the internet. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement* (IMC'04). Association for Computing Machinery, New York, NY, USA, 41-54. DOI:https://doi.org/10.1145/1028788.1028795

Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading.

Tukey, J.W. (1960). *A Survey of Sampling from Contaminated Distributions.* Princeton, New Jersey: Princeton University.

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

Yusof, Z.M., Othman, A.R. and Syed Yahaya, S.S. (2013). Robustness of Trimmed F statistics when handling nonnormal data. *Malaysian Journal of Science*, 32(1): 73-77.