UNIVERSITI PUTRA MALAYSIA

MODIFIED SEQUENTIAL FENCES FOR IDENTIFYING UNIVARIATE OUTLIERS

WONG HUI SHEIN

IPM 2016 21

MODIFIED SEQUENTIAL FENCES FOR IDENTIFYING UNIVARIATE OUTLIERS

By

WONG HUI SHEIN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of theRequirements for the Degree of Master of Science

November 2016

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirements for the degree of Master of Science

MODIFIED SEQUENTIAL FENCES FOR IDENTIFYING UNIVARIATE OUTLIERS

By

WONG HUI SHEIN

November 2016

Chairman: Anwar Fitrianto, PhD
Faculty: Institute for Mathematical Research

The existence of outliers in data set can bring some impacts on statistical data analysis and affect decision making. Thus, it is vital for researcher to identify the outliers. Sequential fences is a graphical method which was proposed by Schewertman and de Silva (2007). Besides its simplicity, this method is also effective in detecting multiple outliers while maintaining the approximate specific outside rate at each stage as the series on number of outlier fences. This research focuses on the modification of sequential fences to improve its efficiency.

Sequential fences method is modified by replacing interquartile range with various robust scales such as semi-interquartile range, Qn, Sn, median absolute deviation (MAD) and Gini's mean difference (GMD) in order to improve outlier detection in symmetric distribution. Ultimately, the utilisation of GMD in sequential fences seems to demonstrate a comparable accuracy in detecting the contaminated data. We have shown that GSF approach effectively reduce the masking and swamping problems in identifying the outliers.

Furthermore, a new approach is proposed by considering the skewness of underlying distribution to increase efficiency of sequential fences in skewed distribution. Conclusively, based on the numerical examples and simulation study, newly proposed method has been adjusted according to the skewness of the underlying distribution of data. The results show that the new approach performed better in reducing swamping effect which is misclassifying non-contaminated observation as outlier in asymmetric distribution.

Moreover, we proposed a new method with modified algorithm and methodology namely bootstrap sequential fences. The proposed method involves initial screening of data and bootstrap technique to improve the performance of sequential fences. The modified sequential fences method is found can accurately detect the outliers in positively skewed distribution. In addition, this proposed method also estimates trimmed mean and trimmed standard deviation with smaller bias and smaller root of mean squares error. Thus, proposed method proves its superiority over the existing techniques.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk Ijazah Master Sains

PAGAR BERURUTAN TERUBAHSUAI BAGI MENGENAL PASTI DATA UNIVARIAT TERPENCIL

Oleh

WONG HUI SHEIN

November 2016

Pengerusi: Anwar Fitrianto, PhD
Fakulti: Institut Penyelidikan Matematik

Kewujudan data terpencil dalam data boleh membawa kesan negatif terhadap analisis data statistik dan menjejaskan kesimpulan. Oleh itu, ini adalah penting bagi penyelidik untuk mengenal pasti data terpencil. Pagar berurutan adalah satu kaedah grafik yang dicadangkan oleh Schewertman dan de Silva (2007). Selain mudah, kaedah ini juga berkesan dalam mengesan pelbagai data terpencil disamping mengekalkan kadar luar tertentu yang sesuai pada setiap peringkat sebagai siri pada bilangan pagar titik terpencil. Kajian ini memberi tumpuan kepada pengubahsuaian pagar berurutan untuk meningkatkan kecekapannya.

Kaedah pagar berurutan telah diubahsuai dengan menggantikan julat antara kuartil dengan pelbagai skala teguh seperti julat semi-antara kuartil, Qn, Sn, sisihan mutlak median (MAD) dan perbezaan min Gini (GMD) untuk meningkatkan pengecaman data terpencil dalam taburan simetri. Penggunaan GMD dalam pagar berurutan menunjukkan ketepatan yang setanding dalam mengesan data yang tercemar. Kami telah menunjukkan bahawa pendekatan GSF berkesan dalam mengurangkan masalah litupan dan limpahan dalam mengenal pasti titik terpencil.

Selain itu, satu pendekatan yang baru telah dikemukakan dengan mempertimbangkan kepencongan taburan dasar untuk meningkatkan kecekapan pagar berurutan dalam taburan pencongan. Kesimpulannya, berdasarkan contoh-contoh berangka dan simulasi kajian, pendekatan baru yang dicadangkan telah disesuaikan mengikut kepencongan taburan pendasar data. Keputusan menunjukkan bahawa pendekatan baru memberikan prestasi yang lebih baik dalam mengurangkan kesan limpahan yang tersilap mengklasifikasikan titik bukan tercemar sebagai titik terpencil dalam taburan bukan simetri.

Di samping itu, kami juga mencadangkan satu kaedah baru dengan algoritma dan kaedah yang diubahsuai iaitu bootstrap pagar berurutan. Kaedah yang dicadangkan melibatkan pemeriksaan awal data dan teknik bootstrap untuk mepertingkatkan prestasi pagar berurutan. Kaedah pagar berurutan yang diubahsuai didapati bahawa boleh mengesan titik terpencil dengan tepat dalam lengkung pencong positif. Tambahan pula, pendekatan baru ini juga menunjukkan kecenderungan dan punca kuasa dua min ralat yang lebih kecil dalam penganggaran min terpangkas dan sisihan piawai terpangkas. Oleh yang demikian, terbuktilah keunggulan pendekatan baru berbanding dengan teknik-teknik yang sedia ada.

ACKNOWLEDGEMENTS

First of all, I would like to express my special appreciation and thanks to my supervisor, Dr. Anwar Fitrianto for his support, encouragement and valuable advice on all matters related to this master thesis. I highly appreciate his constant guidance towards the completion of this master thesis. Besides, I would also like to thank my co-supervisors, Prof. Habshah binti Midi for her constructive comments and motivation in this research.

I would also like to extend my appreciation to the authority of Universiti Putra Malaysia for providing me with good environment and facilities throughout my study, especially towards the accomplishment of this master thesis. An honorable mention goes to my beloved family members for their understandings and support. Words cannot express how grateful I am to them in inspiring me to strive towards my goal.

Last but not least, I would like to express my heartfelt thanks to my friends and my loved one, Tiaw Kah Fook, for their encouragements, supports and blessings for the successful completion of my study at UPM. Besides, I also take this opportunity to express my sincere gratitude to others who directly or indirectly have lent their helping hands in this venture and made my master research duration more memorable.

I certify that a Thesis Examination Committee has met on 8 November 2016 to conduct the final examination of Wong Hui Shein on her thesis entitled "Modified Sequential Fences for Identifying Univariate Outliers" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Master of Science.

Members of the Thesis Examination Committee were as follows:

Jayanthi a/p Arasan, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Noor Akma binti Ibrahim, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Yong Zulina Zubairi, PhD
Associate Professor
University of Malaya
Malaysia
(External Examiner)

NOR AINI AB. SHUKOR, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 27 December 2016

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirements for the degree of Master of Science. The members of the Supervisory Committee were as follows:

Anwar Fitrianto, PhD
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Habshah bt Midi, PhD
Professor
Faculty of Science
University Putra Malaysia
(Member)

____
ROBIAH BINTI YUNUS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, leaning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the University Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature:____________________________ Date: _________________________

Name and Matric No. :Wong Hui Shein, GS41099

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ______________________________
Name of Chairman of Supervisory Committee: Dr. Anwar Fitrianto

Signature: ______________________________
Name of Member of Supervisory Committee: Prof. Dr. Habshah bt Midi

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASF	Adjusted sequential fences
ESD	Generalised extreme studentized deviation test
GCD	procedure that is incorporating the GSF approach into the algorithm to generate a set of clean data
GMD	Gini's mean difference
GSF	Sequential fences using GMD
IQR	interquartile range
LF	Lower fence
MAD	Median absolute deviation
MADSF	Sequential fences using MAD
MS	Moment of skewness
Q1	First quartile
Q3	Third quartile
QnSF	Sequential fences using Qn
RMSE	Root of mean square error
SDSF	Sequential fences proposed by Schewertman and de Silva (2007)
SDSFB	Sequential fences using the proposed bootstrapping techniques
SIQR	semi-interquartile ranges
SIQRSF	Sequential fences using SIQR
SnSF	Sequential fences using Sn
SSFB	Bootstrap technique is used to estimate the cut off points of sequential fences
TB	Tukey's boxplot
TBB	Tukey's boxplot using the proposed bootstrapping techniques
TEB	Trimmed estimators based on bootstrap resampling
UF	Upper fence

CHAPTER 1

INTRODUCTION

1.1 Outlier Definitions

An outlier is an observation that appears discrepant with the other values of the sample. Outliers can also be defined as those observations that look different from other members in the data **(**Beckman & Cook, 1983). Another definition of an outlier is a value which appears inconsistent to the researcher (Iglewicz & Hoaglin, 1993). In other words, inconsistent observations with respect to the remaining data are defined as outliers. An outlier is also defined as an observation which deviates away from the other data values and this outlying observation is suspected that it was created by other mechanism (Hawkins, 1980).

From the historical definitions, these can be illustrated that an outlier is a subjective and post-data concept. Methods for dealing with outliers are applied to the data for checking the existence of the outliers after the contaminated observations are detected via a visual examination of the data (Beckman & Cook, 1983; Grubbs, 1969). In short, an observation that comes from a distribution that is different from that for all the other remaining observations is determined as a contaminated observation.

1.1.1 Causes and Influences of Outliers

The occurrence of outliers in the data set can be caused by mistake in recording or due to the malfunction of measuring instrument. Besides, the existence of discordant observations might be due to the natural variability which comes from the outside of the sample. These outliers may have great influence on the parametric data analyses and resulted in misleading results. During the estimation of parameters, the presence of outliers may cause high errors variance and low power of test (Zimmerman, 1994, 1995, 1998). When there are outliers in the errors, the normality in univariate case and sphericity and multivariate normality become low and lead to type I and type II errors. In linear regression, the effect of outlier is at least distorting the parameter estimation (Osborne & Overbay, 2004).

1.1.2 Swamping and Masking

Outlier identification plays a vital role in statistical inference, data processing and modeling. The presence of outliers might result in biased parameter, poor forecasting and misspecification in modeling (Tsay et al., 2000; Fuller, 1987). There are many literatures on the outliers detection methods. Some methods might classify clean observations as outliers and fail to detect the real outliers. Thus, swamping and masking effects emerge. The swamping and masking effects can cause mistake in making decision during the regression analysis. (Chatterjee & Hadi, 2006). The characteristics of these effects are defined by Iglewicz and Martinez (1982) and Ben-Gal (2009).

Swamping effect occurs when a second observation is labeled as an outlier in the presence of first outlier. After discarding the first outlying observation, the second observation is detected as clean observation. This phenomenon is classified as the swamping effect. Swamping effect happens when outlier shifts the mean and the covariance estimates toward it and away from other inliers on another side of distribution tail. Hence, this causes the gap between these observations to the mean is large and make them look similar to outliers.

Another phenomenon is that the second observation is classified as outlier without the existence of the first outlier. After eliminating the first outlier, the second observation is appeared as an outlier. This occurrence is denoted as the masking effect. Masking effect occurs when mean and covariance estimates are skewed towards a group of outliers, and the resulting distance of the outlier from the mean is decreased.

1.2 Tukey's Boxplot

Traditional boxplot is one of the most frequently and widely used techniques for studying the shape of the distribution and analyzing some characteristics of the distribution such as location and spread. In addition, the boxplot technique also can be used to identify the potential outliers which deviate markedly from the remaining data.

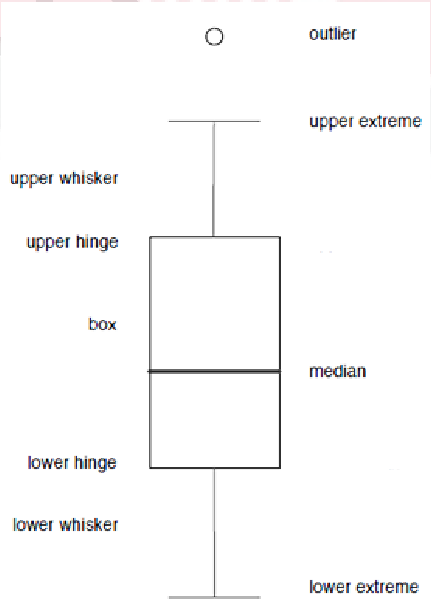

Figure 1.1: Construction of boxplot

The boxplot consists of five components which give a robust statistical summary of the distribution of a dataset. These components are illustrated in Figure 1.1 The components used to construct the boxplot are two hinges which are first quartile (q_1) and third quartile (q_3), median, observations which lie 1.5 constant from the interquartile range measured from median, two whiskers that connect to the lower and upper hinges and potential outliers which lie apart from median and exceed extreme. The first quartile and third quartile are equivalent to 25^{th} percentile and 75^{th} percentile respectively while the interquartile range is the difference between the third and first quartile.

Inner fences in a boxplot are positioned at an interval of 1.5 IQR beneath first quartile and above third quartile which can be denoted as

$$[q_1 - 1.5\text{IQR}, q_3 + 1.5\text{IQR}] \tag{1.1}$$

whereas the outer fences are located at a distance of 3 IQR less than first quartile and more than third quartile which are presented as

$$[q_1 - 3\text{IQR}, q_3 + 3\text{IQR}]. \tag{1.2}$$

Any observation that falls outside the inner fences is labeled as a mild outlier while value that falls beyond outer fences is marked as extreme outlier.

1.3 Sequential Fences

Schewertman and de Silva (2007) has modified the boxplot and introduced sequential fences as another useful technique to detect the outliers in the data. In this study, the sequential fences proposed by Schewertman and de Silva is henceforth referred as SDSF. The technique proposed by Schewertman and de Silva (2007) identifies outliers sequentially based on the specific sample size and the pre-specified outside rate attained which is the probability that an uncontaminated observation falls beyond the fences.

For the construction of sequential fences, the sample sizes are adjusted using Poisson model in order to decrease the tail probabilities. The adjustment is similar to the adjustment done in Davies and Gather (1993) and Gather and Becker (1997). This SDSF increases the accuracy to identify the outliers, reduces the swamping effect and less likely to misclassify an uncontaminated observation as an outlier in large sample size. In the procedure of outlier identification, this method allows the researchers to have flexibility in setting the level of confidence. The fences are constructed continuously until there is no extra outlier detected.

1.4 Problem Statement

Although there are a lot of literatures on outlier identification methods, most of the existing methods are suitable only for symmetric distributions as discussed in detail in Chapter 2. The popular boxplot method (Tukey, 1977) is too liberal and cause many unusual observations to be overlooked. The sequential fences method which was proposed by Schwertman and de Silva (2007) allows flexibility in setting the outside rate to detect the extreme and mild outliers. This method uses interquartile range to measure the dispersion of the data. A natural question comes to our mind is whether it can be developed using an alternative robust scale that can measure the dispersion of the data in the sequential fences method. Thus, it is important to find out the suitable robust scale in the replacement of interquartile range in order to improve the performance of sequential fences approach in detecting the outliers.

Furthermore, the major problem of the existing outlier detection techniques is too conservative in which these techniques work well in symmetric distribution and have low performance in asymmetric distribution. Some methods obey normality assumptions while most of the real data do not follow normal distribution. Some authors proposed outliers techniques for skewed data, but the performance of these techniques needs improvement. Therefore, the modification of the sequential fences method which was proposed by Schwertman and de Silva (2007) is needed to be improved by making some adjustments to the approach for detecting outliers in skewed data with the consideration of the skewness of the distributions.

Moreover, procedure of screening for the data before further analysis of data is important (Tabachnick & Fidell, 2001). Identification of outliers is a part of the data screening procedure which should be done regularly before starting a statistical analysis (Beckman & Cook, 1983; Ahmad et al., 2011). Simulated univariate data may contain outlying observations. When the data is from symmetric distribution, the extreme values that are located at the left or right tail may be suspected as outliers. For skewed distribution data, it is suspected that the extreme observation at the longer tail might be outlying observation.

In order to know whether the outliers present in the data, initial screening of the data is necessary. In the boxplot method, the data which are used to obtain the central tendency and spread of data such as mean and standard deviation are assumed normal. Test statistics are greatly affected when the data is non-normal. The critical values of sequential fences technique (Schwertman & de Silva, 2007) depends on the calculation of median and interquartile range. In Chapter 4, it can be observed that the existing sequential fences perform well in the symmetric distributions but capture too much outliers on the long tail of skewed distribution. Thus, the fences should be adjusted to allow a better coverage of the centre of the data especially when the data are skewed.

1.5 Objectives of Study

Since the study is focused on modification of sequential fences, the objectives of this study are i) to propose method for outlier identification in symmetric distribution with higher accuracy and lower misclassification of non-contaminated observations as outliers ii) to increase the accuracy in detecting the real outliers in asymmetric distribution with minimum swamping and masking effect and iii) to provide an efficient sequential fences in identifying outliers in wider types of distributions with new algorithm and parameters estimation.

1.6 Limitation of Study

SAS review 9.3 is selected as our research tool which helps in simulation, bootstrapping and computing the results. Due to the long computation time in large replications, the number of observations contamination is set up to three outliers. The sample size of the simulation is limited to $n = 100$ only, because simulations involve 10,000 replications. The procedures of sequential fences take some time because this method has to keep constructing the fences sequentially and checking for the presence of outliers continuously until there is no additional outlying observation being captured.

1.7 Overview of Thesis

Since this study is related to modification of existing sequential fences method (SDSF) which was proposed by Schwertman and de Silva (2007), it is important to improve its performance in outlier detection in symmetric and asymmetric distribution data.

Chapter 3 provides a review of sequential fences method of detecting the outliers in the normally distributed data. Instead of using interquartile range (IQR), this study modifies the existing sequential fences technique for identifying outliers by using different robust scales such as semi-interquartile range (SIQR), median absolute deviation (MAD), Qn, Sn and gini's mean difference (GMD). This study also compares proposed methods with the existing sequential fences method and generalized extreme studentized deviate (ESD) test. Two empirical examples are used to illustrate the efficiency of the methods. Simulation study is conducted with different number of outliers. The performance of all outlier detection techniques has been compared by evaluating the proportion of correctly identifies the outliers and the proportion of misclassifies the uncontaminated observation as outliers. Superiority of the proposed technique has been validated by simulation results.

In Chapter 4, SDSF method is extended to form a new technique based on the skewness of underlying distribution data to identify outliers in skewed distributions. Adjustment of the fences construction has been made using moment measure of skewness to measure the skewness of the data. Similarly, the proposed method and existing SDSF are applied to a real data set as illustration and comparison. Besides Normal distribution, simulation study has been conducted on skewed distributions such as Lognormal, Chi-square, Gamma and Weibull with different parameters, and the simulation results are compared with existing SDSF method. The proposed technique shows its outstanding performance compared to SDSF technique in detecting outliers in the different distributions and also in real data set at different nominal outside rates.

In Chapter 5, a modification of algorithm and formulation are proposed based on the SDSF method which can identify outliers in the symmetric and asymmetric distributions. Instead of using Monte Carlo simulation, a new methodology involving bootstrapping technique has been developed. Before contamination of the data, a clean simulated data is generated and verified using Gini Sequential Fences (GSF) method which is proposed in Chapter 3. For the performance study, bootstrap resampling study has been done on the symmetric and skewed distribution, such as normal, chi-square with different degrees of freedom and lognormal distribution with different parameters. The performance of the newly proposed technique is compared with SDSF method and Tukey's boxplot by matching number of outliers detected with the contaminated observations for different sample sizes. Apart from that, based on the outliers detected, trimmed mean and trimmed standard deviation adopting bootstrap resampling technique has been calculated. The comparison of the estimation of parameters based on bias and mean square errors have been done. From the result, the supremacy of proposed modification method over existing SDSF technique and boxplot is proven.

REFERENCES

Adil, I. H., & Irshad, A. U. (2015). A Modified Approach for Detection of Outliers. *Pakistan Journal of Statistics and Operation Research*, *11*(1), 91.

Ahmad, S., Midi, H., & Norazan, M. R. (2011). Diagnostics for Residual Outliers Using Deviance Component in Binary Logistic Regression, *World Applied Sciences Journal*, *14*(8), 1125-1130.

Akerlof, C. (1983). Efficient algorithms for estimating the width of nearly normal distributions. *Nuclear Instruments and Methods in Physics Research*, *211*(2-3), 439-445.

Aleskerov, E., Freisleben, B., and Rao, B. (1997). Cardwatch: A neural network based database mining system for credit card fraud detection. In Proceedings of IEEE Computational Intelligence for Financial Engineering. 220-226.

Fitrianto, A. & Midi, H. (2011). Procedures of Generating a True Clean Data in Simple Mediation Analysis. *World Applied Sciences Journal*, *15*(7), 1046-1053.

Armin B. (2008). One-sided and Two sided Critical Values for Dixon's Outlier Test for Sample Sizes up to $n = 30$, *Economic Quality Control*, *23*(1), 5-13.

Aslam M. & Khurshid A. (1991). Shape-finder box plots. *ASQC Statistics Division Newsletter*, 9–11.

Aucremanne, L., Brys, G., Hubert, M., Rousseeuw, P. J., & Struyf, A. (2004). A Study of Belgian Inflation, Relative Prices and Nominal Rigidities using New Robust Measures of Skewness and Tail Weight. *Theory and Applications of Recent Robust Methods*, 13-25. doi:10.1007/978-3- 0348-7958-3_2

Barbato, G., Barini, E. M., Genta, G., & Levi, R. (2011). Features and performance of some outlier detection methods, *Journal of Applied Statistics*, 38:10, 2133-2149.

Barnett, F. C., Mullen, K., & Saw, J. G. (1967). Linear estimates of a population scale parameter. *Biometrika*, *54*(3-4), 551-554.

Barnett, V. (1978). The Study of Outliers: Purpose and Model. *Applied Statistics, 27* (3), 242-250.

Barnett, V., Lewis, T. (1994). *Outliers in Statistical Data.3rd Ed*. Chichester: John Wiley.

Beckman. R.J. and R.D. Cook. (1983). Outlier….s, *Technometrics*, 25, 119-149.

Bendre, S. M., & Kale, B. K. (1987). Masking effect on test for outliers in normal sample. *Biometrika , 74* (4), 891-896.

Ben-Gal, I. (2009). Outlier Detection. *Data Mining and Knowledge Discovery Handbook*, 117-130. doi:10.1007/978-0-387-09823-4.

Boos, D. D., & Stefanski, L. A. (2013). *Essential statistical inference: Theory and methods*.

Brant, R. (1990). Comparing Classical and Resistant Outlier Rules. *Journal of the American Statistical Association*, *85*(412), 1083-1090.

Cao, D. S., Liang, Y. Z., Xu, Q. S., Li, H. D., Chen, X. (2010). A new strategy of outlier detection for QSAR/QSPR. *Journal of Computational Chemistry* 31(3):559–602

Capéràa, P., & Rivest, L. P. (1995). On the variance of the trimmed mean. *Statistics probability letters*, *22*(1), 79-85.

Carling, K. (2000).Resistant outlier rules and the non-Gaussian case.*Computational Statistics and Data Analysis*, *33*(3), 249-258.

Cerioli, A. (2010). Outlier detection with high-breakdown estimators. *Journal of the American Statistical Association* 105:147–156.

Chatterjee, S., & Hadi, A. S. (2006). *Regression Analysis by Example* (Fourth Edition ed.). Hoboken, New Jersey: John Wiley and Sons, Inc.

Choonpradub, C. & McNeil, D. (2005). Can the box plot be improved? *Songklanakarin Journal of Science and Technology*, *27*(3), 649–657.

Dang, X., Serfling, R. (2010).Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *Journal of Statistical Planning and Inference* 140:198–213.

David, H. A.(1968). Gini's Mean Difference Rediscovered. *Biometrika*, *55*(3), 573–575.

Davies, L., Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, *88*, 782-792.

Dixon, W. J. (1950). Analysis of Extreme Values. *Annals of Maths. Stat. 21*, 488-506.

Doane, D. P., & Seward, L. E. (2011). Measuring skewness: A forgotten statistic? *Journal of Statistics Education*, 19(2), 1-18.

Dovoedo, Y. H., & Chakraborti, S. (2012). Boxplot-based Phase I control charts for time between events. *Quality and Reliability Engineering International 28*(1):123-130.

Dovoedo, Y. H., & Chakraborti, S. (2013). Outlier detection for multivariate skew-normal data: A comparative study. *Journal of Statistical Computation and Simulation* 83(4):773–783.

Dovoedo Y. H., & Chakraborti, S. (2015). Boxplot-Based Outlier Detection for the Location-Scale Family. *Communications in Statistics – Simulation and Computation*, *44*(6), 1492-1513.

Downton, F. (1966). Linear estimates with polynomial coefficients. *Biometrika*, *53*, 129-41.

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

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1-26. doi:10.1214/aos/1176344552

Ferguson, T.S. (1961). On the rejection of outliers. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1: 377-381.

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some Implementations of the Boxplot. *The American Statistician*, *43*(1), 50-54.

Fuller, W. A. (1987). *Measurement Error Models.* United States of America: Braun-Brumfield, lnc.

Gather, U., & Becker, C. (1997). Outlier identification and robust methods. *Handbook of Statistics Robust Inference*, *15*, 123-143.

Gerstenkorn, T., & Gerstenkorn, J. (2003). Gini's mean difference in the theory and application to inflated distributions, *Statistica*, *63*(3), 469-488.

Gerstenberger, C., & Vogel, D. (2015). On the efficiency of Gini's mean difference. *Statistical Methods & Applications*, *24*(4), 569-596.

Ghosh, S. and Reilly, D. L. (1994). Credit card fraud detection with a neural-network. *Proceedings of the 27th Annual Hawaii International Conference on System Science HICSS-94*. doi:10.1109/hicss.1994.323314

Gini C. (1912).Variabili à e mutabilita,contributoallo studio delledistribuzioni e relazionistatistiche, *StudiEconomico – Giuridicidella R. Universita di Cagliari*, *3*(2), 3-159.

Gross, A. (1976). Confidence Interval Robustness with Long-Tailed Symmetric Distribution, *Journal of the American Statistical Association*, *71*(354), 409-416.

Grubbs, F.E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*, 21: 27-58.

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, *11* (1), 1-21.

Hawkins, D. M. (2006). Outliers. *Encyclopedia of Statistical Sciences*. New York: John Wiley.

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-Tuning Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association, 82*(400), 1147-1149.

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association, 81*(396), 991-999.

Hubert, M., & Vandervieran, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis, 52*(12), 5186-5201.

Hyndman, R. J., & Fan, Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistician, 50*(4), 361.

Iftikhar, H. A. (2011). Robust outlier detection techniques for skewed distributions and applications to real data (Doctoral thesis, International Islamic University, Islamabad, Pakistan).

Iglewicz, B., & Hoaglin, D. C. (1993). How to Detect and Handle Outlier. 16, Wisconsin: ASQC Quality Press.

Iglewicz, B., & Martinez, J. (1982). Outlier detection using robust measures of scale. *Journal of Statistical Computation and Simulation*, *15*(4), 285-293. doi:10.1080/00949658208810590

Iglewicz. B.,& Banerjee, S. (2001). A simple univariate outlier identification procedure. Proceedings of the American Statistical Association.

Irwin, J.O. (1925). On a criterion for the rejection of outlying observations. *Biometrika*, 17: 238-250.

Kendall, M. G., & Stuart, A. (1958). The advanced Theory of Statistics. 1. London: Griffin.

Kimber, A. C. (1990). Exploratory Data Analysis for Possibly Censored Data From Skewed Distributions. *Applied Statistics, 39*(1), 21-30.

Kumar, V. (2005). Parallel and Distributed Computing for Cybersecurity. Distributed Systems Online, IEEE 6, 10.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement*, *58*(3), 409- 429.

Louni, H. (2008). Outlier detection in ARMA models. *Journal of Time Series Analysis 29*(6):1057–1065.

Marmolejo, R. F. & Tian, T. S. (2010). The shifting boxplot: A boxplot based on essential summary statistics around the mean. *International Journal of Psychological Research 3*(1), 37–45.

Montgomery, D. C. (2009). *Introduction to Statistical Quality Control*. 6th ed. New York: John Wiley.

Montgomery, D. C., Peck, E. A., Vining, G. G. (2001). *Introduction to Linear Regression Analysis*. 3rd ed. New York: John Wiley.

Nelson, A. C, Armentrout, D. W. & Johnson, T. R.. (1980). Validation of Air Monitoring Data, EPA- 600/4-80-030. U. S. Environmental Protection Agency, Research Triangle Park, N. C.

Rosner, B. (1975). On the detection of many outliers. *Technometrics*, 17: 221-227.

Rosner, B. (1983). Percentage Points for a Generalized ESD Many-Outlier Procedure, *Technometrics*, *25*(2), 165.

Rousseeuw P. J., & Croux, C. (1993). *Journal of the American Statistical Association*, *88*(424), 1273-1283.

Rousseeuw, P. J., & Leroy, A. M. (1987). Robust Regression and Outlier Detection. *Wiley*, New York.

Schwertman, N. C., de Silva, R. (2007). Identifying outliers with sequential fences. *Computational Statistics and Data Analysis* 51:3800–3810.

Schwertman, N. C., Owens, M. A., Adnan, R. (2004). A simple more general boxplot method for identifying outliers. *Computational Statistics and Data Analysis* 47:165–174.

Seo, S. (2006). A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. University of Pittsburgh, Graduate School of Public Health.

Singh, K., & Xie, M. (2003). Bootlier-Plot: Bootstrap Based Outlier Detection Plot. *Sankhyā: The Indian Journal of Statistics (2003-2007)*, *65*(3), 532-559. Retrieved from http://www.jstor.org/stable/25053287

Spence, C., Parra, L., and Sajda, P. (2001). Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. IEEE Computer Society, Washington, DC, USA, 3.

Stigler, S. M. (1977). Do Robust estimators work with real data. *The Annals of Statistics*, *5*(6), 1055-1098.

Tabachnick, B. G. & Fidell, L. S. (2001). *Using Multivariate Statistics, fourth edition.* New York: Boston and Bacon.

Tsay, R. S., Pena, D., & Pankratz, A. E. (2000). Outliers in Multivariate Time Series. *Biometrika* , *87* (4), 789-804.

Tukey, J. W. (1977). *Exploratory Data Analysis*. New York: Addison-Wesley.

Walsh, J. (1959). Large sample nonparametric rejection of outlying observations. *Annals of the Institute of Statistical Mathematics 10*, 223–232.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research, 65*(1), 51-77.

Yitzhaki, S.(2003). Gini's Mean difference: A superior measure of variability for non-normal distributions, *Metron*, *61*(2),285-316.

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology* , *121*(4), 391-401.

Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. *Journal of Experimental Education* , *64*(1), 71-78.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education* , *67*(1), 55-68.

BIODATA OF STUDENT

Wong Hui Shein was born on 29th March 1990 in Ipoh, Perak. She began her primary school education at SJK (c) Chin Hua and received her secondary education at SMK Dato' Zulkifli Muhammad. She then continued her pre-university education (Form 6) in SMK Khir Johari. After that, she enrolled in Universiti Putra Malaysia in the program of Bachelor of Science (Honours) Majoring in Statistics and graduated with 1st Class Honours. In September 2014, she furthered her study at the same university in the field of Applied and Computational Statistics.

LIST OF PUBLICATIONS

Wong, H. S., & Fitrianto, A. (2016) A Comparative Study of Outliers Identification Methods in Univariate Data Set. *Journal of Advanced Science Letters.*

Wong, H. S., & Fitrianto, A. (2016) Outliers Detection using Sequential Fences with different Robust Scales. *Communications in Statistics - Simulation and Computation.*

UNIVERSITI PUTRA MALAYSIA

STATUS CONFIRMATION FOR THESIS / PROJECT REPORT AND COPYRIGHT

ACADEMIC SESSION : ____________________

TITLE OF THESIS / PROJECT REPORT :

MODIFIED SEQUENTIAL FENCES FOR IDENTIFYING UNIVARIATE OUTLIERS

NAME OF STUDENT: WONG HUI SHEIN

I acknowledge that the copyright and other intellectual property in the thesis/project report belonged to Universiti Putra Malaysia and I agree to allow this thesis/project report to be placed at the library under the following terms:

1. This thesis/project report is the property of Universiti Putra Malaysia.
2. The library of Universiti Putra Malaysia has the right to make copies for educational purposes only.
3. The library of Universiti Putra Malaysia is allowed to make copies of this thesis for academic exchange.

I declare that this thesis is classified as :

*Please tick (√)

☐	CONFIDENTIAL	(Contain confidential information under Official Secret Act 1972).
☐	RESTRICTED	(Contains restricted information as specified by the organization/institution where research was done).
☐	OPEN ACCESS	I agree that my thesis/project report to be published as hard copy or online open access.

This thesis is submitted for :

☐	PATENT	Embargo from _____________ (date) until _____________ (date)

Approved by:

____________________	____________________
(Signature of Student)	(Signature of Chairman of Supervisory Committee)
New IC No/ Passport No.:	Name:
Date :	Date :

[Note : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization/institution with period and reasons for confidentially or restricted.]