

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِیْمِ

قال تعالى :

وَبِرَّحْمَتِهِ لَبِئْسَ الَّذِيْنَ كَفَرُوْا ۗ وَهُوَ خَيْرٌ
مِّمَّا يَجْمَعُوْنَ (58))

سورة يونس الايه (58)

Dedicated to

The memory of my father.

My most dear mother.

My wife.

My brothers and sisters.

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor Porf. Z. A.

EL- Beshir for his guidance and advice through out my research.

Thanks are also due to my family for the support and encouragement I

Received form it.

Mr. Fakhr eldin Ismail also deserve thanks for his valuable help in

Writing and executing the necessary computer programmers.

Finally my thanks extend to all those friends and colleagues who

Provided helps of various kinds.

المستخلص

يهدف هذا البحث لدراسة مشكلة اكتشاف القيم الشاذة في البيانات ذات البعد الواحد عندما يكون توزيعها ملتويا لليمين .

إقترحت في الأطروحة طريقه لامعلميه جديده تتضمن قاعده محافظه نسبيا لاكتشاف القيم الشاذة.

تمت مقارنة أداء الطريقه المقترحه من خلال دراسة محاكاه مع أربعة من الطرق اللا معلميه شائعة الاستخدام. وقد أجريت المقارنه بإفتراض ثلاثة توزيعات للبيانات وهي توزيعات مربع كاي وبيتا والتوزيع الطبيعي اللوغاريتم .

وقد استخدمت أحجام عينات صغيره ومتوسطه وكبيره في التجريه.

وقد أكدت دراسة المحاكاه تفوق الطريقه المقترحه على الطرق الأخرى عند محاولة اكتشاف القيم الشاذة في الطرف الايمن من التوزيع .

ويتزايد هذا التفوق مع زيادة حجم العينه , أما أدائها في الطرف الأيسر فهو ضعيف نسبيا.

Abstract

This research aimed at investigating the problem of outlier detection in univariate data when the distribution is skewed to the right .

A new nonparametric method is proposed that involves a more conservative rule of detection. The performance of the suggested method is compared through a simulation experiment, to that of four widely used nonparametric methods.

The comparison is made under three distributions , namely the chi-square , beta and log normal distribution. Small, medium and large sample size are used in the experiment.

The result of the simulation, confirmed the superiority of the Proposed method to the other methods when testing for outliers in the right tail of the distribution .

This superiority increase with increase in sample size. The Performance of the method in the left tail is relatively poor.

Contents

Dedication to

Acknowledgment

Abstract (Arabic)

Abstract (English)

List of Tables

List of Symbols

1 Introduction

1.1 Statement of the problem.....	1
1.2 Research objective.....	2
1.3 Research Approach.....	2
1.4 Structure of the Thesis.....	2

2 Basic concepts 4

2.1 Introduction.....	4
2.2 Definition of outlier.....	4
2.3 Masking and Swamping effect	5
2.4 Type of outliers.....	6
2.5 Source of outliers.....	7
2.6 Importance of Detecting Outliers.....	11
2.7 Applications of Outlier Detecting Techniques	12
2.8 Skewness.....	13
2.9 Breakdown point.....	14

3 Literature review.....	15
3.1 Introduction.....	15
3.2 A historical review of Detection methods of Uniavarit Outliers.....	15
3.3 Influence Measures.....	22
3.4 Various methods for identification of outliers.....	25
3.4.1 Test of discordance.....	26
3.5 Outlier labeling methods.....	28
3.5.1 Univariate outlier detection.....	29
3.5.2 Univariate Robust methods.....	31
3.6 Multivariate method.....	36
3.6.1 Statistical methods.....	36
3.6.2 Multivariate robust measures.....	38
4 A proposed Method.....	41
4.1 Introduction.....	41
4.2 Tukey's methods.....	42
4.3 The proposed method.....	43
4.4 Methodology.....	44
5 Compares of Method.....	46
5.1 Introduction.....	46
5.2 Chi square distribution.....	47
5.3 Lognormal Distribution	55
5.4 Beta Distribution.....	62
6 Conclusion.....	70
6.1 Main Results.....	70
6.2 Recommendation for future work.....	71

Bibliography	72
APPENDIX	79

List of Tables

Table 3.1 Dixon tests for univariate normal samples	28
Table 5.1 Fences of five methods and True boundary in χ^2 Distribution at Small sample size (n=25).....	48
Table 5.2 Fences of the five methods and True boundary in χ^2 Distribution at meduieim sample size (n=100).....	51
Table 5.3 Fences of the five methods and True boundary in χ^2 Distribution at large sample size (n=500).....	53
Table 5.4 Fences of the five methods and True boundary in lognormal Distribution at small sample size (n=25).....	56
Table 5.5 Fences of the five methods and True boundary in lognormal Distribution at medium sample size (n=100).....	58
Table 5.6 Fences of the five methods and True boundary in lognormal Distribution at large sample size (n=500).....	60
Table 5.7 Fences of the five methods and True boundary in β Distribution at Small sample size (n=25).....	63
Table 5.8 Fences of the five methods and True boundary in β Distribution at medium sample size (n=100).....	65
Table 5.9 Fences of the five methods and True boundary in β Distribution at large sample size (n=500).....	67

List of Symbols

MD^2 = Mahalanobis distance

CD^2 = Cook's squared distance

$DFFIT_{S(i)}$ = The number of standard deviations from the i^{th} observation

$DFBETAS$ = A diagnostic measure based on the change in the j^{th} regression coefficient of $\beta_{j(i)}$.

COVRATIO's = Determinant of a covariance matrix of the estimates after deleting the i^{th} observation

AIC = Applied Akaike Criterion

h_{ii} is the i^{th} diagonal element of $X'(X'X)^{-1}X$

\bar{x} = Sample mean

SD = Standard deviation

G = **Grubs test**

r = **Dixon test**

R = **Extreme Studentized Deviate**

MAD = Median Absolute Deviation

$MAD_e = 1.483 \times MAD$

GLDs = Generalized Lambda Distributions

$Z_{i=}$ **Z-SCORE**

Q_1 = Lower quartile

Q_2 = Median

Q_3 = Upper quartile

IQR= Inter quartile range

MC= Medcouple

L= Lower critical value

U= Upper critical value

MCD= Minimum covariance determinant

MVE= Minimum volume ellipsoid

M=M-estimator

$|X_{LL}|$ = Largest absolute value

$|X_{SL}|$ = Smallest absolute value

b= Difference between $|X_{LL}|$ and $|X_{SL}|$