Vol. 14. Issue.2. 2026 (April-June)

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

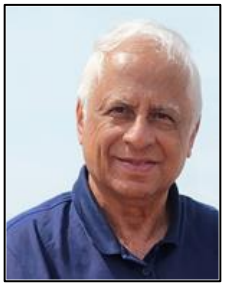
http://www.bomsr.com
Email:editorbomsr@gmail.com

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

RESEARCH ARTICLE

Methods for Detecting the Outliers in Large Samples

Dr. Ramnath Takiar
Scientist G (Retired), National Centre for Disease Informatics and Research (NCDIR), Indian Council of Medical Research (1978–2013), Bengaluru – 562110, Karnataka, India & Flat No. 16, 7th Floor, Building #27B, 1st Khoroo, Ulaanbaatar District, Ulaanbaatar, Mongolia – 14230
Email: ramnath_takiar@yahoo.co.in; ramnathtakiar@gmail.com
DOI:10.33329/bomsr.14.2.1


Dr. Ramnath Takiar

Article Info

Article Received: 13/03/2026
Article Accepted: 10/04/2026
Published online: 23/04/2026

Abstract

Outliers are those observations which lie at an abnormal distance from other observations. The presence of Outliers in data may often result in a skewed distribution, altered kurtosis, inflated mean and Standard deviation. In any statistical data analysis, the presence of Outliers often pose a problem.

For the present study, all observations lying outside the range of a sample are taken as outliers. This is almost equivalent to claim that all observations lying beyond Mean±3SD are outliers. In a sample, the minimum and maximum numbers are replaced by still lower and higher value and treated as outliers. In the present study, overall 200 outliers are introduced, spread over four samples, five sample size and five trials. The five sample sizes chosen are 40, 60, 100, 200 and 300.

For the present study, three method are used for the detection of the Outliers. According to selected three methods to identify the Outliers, the Lower Fence (LF) and Higher Fence values are defined as follows: IQR-Old Method: LF = Q_1 –1.5*IQR and HF= Q_3+ 1.5* IQR. IQR-Takiar method: LF=Q1 – IQR*[0.25*ln(n)+0.20] and HF= Q3+ IQR* [0.25*ln(n)+0.20]. SD-Range-Takiar method: LF= Mean- SD*(0.37*ln (n) + 0.86) and HF = Mean + SD*(0.37*ln (n) + 0.86) where n is the sample size.

Out of 200 outliers introduced, IQR-Old method could detect only 91(45.5%) outliers. SD-Range Takiar method could detect 69 (34.5%)

of the outliers, while IQR-Takiar method could detect 122 (61.0%) of the outliers, correctly. Based on the study results, for large samples, the IQR-Takiar method is adjudged to be the superior method as compared to other two methods in detection of the Outliers. Further, IQR-Takiar method is recommended for detection of outliers in large samples.

Key Words: IQR-Old method, IQR-Takiar method, SD-Range Takiar method, Large samples, Outliers, Outlier detection rate.

Introduction

In any statistical data analysis, it is imperative to judge whether the data collected is appropriate. In such analysis, the presence of Outliers often poses a problem. Outliers are those observations which do not fit into the general pattern of data and behave quite differently with respect to other observations. They visibly lies at an abnormal distance from other observations. The presence of Outliers in data may often result in a skewed distribution, altered kurtosis, inflated mean and Standard deviation. Sometimes, in the presence of Outliers, the relationship studied between any two variables like income and expenditure or age and mortality may present a distorted form of relationship between them(Takiar R, 2026). There are studies available which dealt with the problem of Outliers in different context (Onoz & Oguz 2003, Cousineau & Chartier 2010, Hansen et al. 2023, Lacobucci et al. 2025).

Exploring the relationship between the SD and the range, Takiar (Takiar 2023a) demonstrated that their relationship can be characterized by the equation; Range = SD*(0.73*ln(n) + 1.72). Assuming further that the Range is equally distributed about the mean, for detection of outliers, the Low fence and High fence values are defined as LF = Mean - SD*(0.37*ln (n) + 0.86) and HF = Mean + SD*(0.37*ln (n) + 0.86). In the same year, based on IQR (Takiar 2023b), the two fence values for identification of outliers are defined as LF = Q_1-IQR*(0.25*ln (n) + 0.20) and HF = Q_3 + IQR*(0.25*ln (n) + 0.20). In both the studies, the ability of the methods in detection of outliers were demonstrated on selected samples of size 20 and compared with that of results obtained with IQR-Old method. It was noted that each method gave different set of outliers and there was no way to decide which method is giving the true set of outliers.

In practice we come across only samples and very rarely the population parameters are known. In the absence of knowledge about population parameters, they are estimated from the sample and assumed to represent the population parameters. According to theory, Mean±3SD covers around 99% of the observations. By choosing the Range of the sample, we are covering all the 100% of the observations. Thus, any observations outside the sample range can be considered as the Outliers bringing more objectivity in the definition of the Outliers (Takiar 2026). He evaluated three methods namely IQR-Old, IQR-Takiar and the SD-Range-Takiar method to detect the outliers from the set of four samples each of the sample size 15, 20 and 25, modified artificially by the introduction of the outliers in ten trials in a systematic manner. The Percentage detection rate of Outliers by the IQR-Old method, SD-Range-Takiar method and the IQR-Takiar method was noted to be 42%, 80% and 100%, respectively. Accordingly, it was concluded that the IQR-Takiar method is the superior method in detecting

the outliers and therefore recommended to be used for the identification of Outliers in samples. However, in that study, the evaluations of the methods in detection of Outliers was demonstrated for small samples below 30. In the present study the same three methods are evaluated for detection of outliers in large samples.

Materials and methods

Generation of samples according to different Sample size

For the study purposes, four samples each of size 40, 60, 100, 200 and 300 are generated with predefined mean and the SD, using the function "Generation of Random Numbers" available on the Excel. Thus, 20 random samples are considered for the present study.

Definition of Outlier

For the present study, all the observations lying outside the range of a sample are taken as outliers.

Generation of Outliers

In a sample, the minimum and the maximum numbers are replaced by the still lower and higher values and treated as the Outliers. If the Minimum is denoted by M then it is decreased by the 3% of Mean units each time. Accordingly, the maximum denoted by K is increased by the 3% of Mean units each time. In case, 3% of the Mean comes out to be a fraction number then it was rounded off to the nearest integer as shown in Table 1. In total, five such trials are made resulting in 10 outliers for a sample.

Scheme of Generation of Outliers by Sample and Sample size

The number of outliers generated by sample size and the trial are shown in Table 2. It is to be noted that for each sample size 40 outliers are generated.

Table 1: Generation of Outliers for Each Sample Size

	Sample size				
Variable	40	60	100	200	300
Mean	55.5	65.5	75.5	105	220
SD	16.2	20.5	23.5	32	72
3% of Mean	1.665	1.965	2.265	3.15	6.6
Round off	2	2	2	3	7

Table 2: Generation of Outliers for Selected four Samples

Trial	Sample size				
	40	60	100	200	300
1	8	8	8	8	8
2	8	8	8	8	8
3	8	8	8	8	8
4	8	8	8	8	8
5	8	8	8	8	8
Total	40	40	40	40	40

The following three methods are employed to identify the introduced Outliers :

IQR-Old Method, IQR-Takiar Method and the SD-Range-Takiar Method. For each method, two fence values namely the Lower Fence value (LF) and the Higher Fence value (HF) are defined and shown in Table 3. It is to be noted that these formulae essentially defines the lowest and highest value in a range of the observations for a given sample.

Table 3: Formulae for defining the Lower and Higher Fence values by different methods

Method	Fence value	Formula for defining the Fence value
IQR-Old	LF	$Q_1 - 1.5*IQR$
	HF	$Q_3 + 1.5*IQR$
IQR-Takiar	LF	$Q_1 - IQR*[0.25*\ln(n)+0.20]$
	HF	$Q_3 + IQR*[0.25*\ln(n)+0.20]$
SD-Range -Takiar	LF	Mean - SD*(0.37*ln (n) + 0.86)
	HF	Mean + SD*(0.37*ln (n) + 0.86)

Results

The details of the selected samples like Mean, SD, Skewness and Kurtosis are provided by the sample size in Table 4.

Table 4: The Details of the Selected Samples by Sample size

Sample Size	Sample	Mean	SD	Min	Max	Skewness	Kurtosis
40	1	54.9	13.43	21.5	73.5	-0.59	-0.33
	2	61.0	19.74	23.1	99.2	0.17	-0.59
	3	54.2	16.1	25.9	98.6	0.52	0.2
	4	52.8	16.46	20.7	85.8	0.05	-0.95
60	1	61.1	22.65	9.9	111.1	-0.17	-0.32
	2	64.1	16.38	28.3	99.0	-0.05	-0.52
	3	66.8	18.2	30.8	112.2	0.17	-0.4
	4	62.6	21.4	17.7	124.2	0.34	0.47
100	1	75.7	23.79	26.9	150.3	0.47	0.45
	2	77.1	23.24	35.5	143.4	0.52	0.13
	3	74.2	23.31	21.5	130.3	0.25	-0.39
	4	75.3	19.06	23.2	126.0	0.19	-0.06
200	1	108.4	36.6	2.4	206.4	-0.05	-0.06
	2	103.7	31.64	33.7	206.4	0.24	-0.03
	3	108.7	32.26	26.0	196.4	0.08	-0.21
	4	107.5	31.75	25.9	233.3	0.35	0.64

300	1	212.9	71.95	8.2	453.9	-0.02	-0.01
	2	218.7	69.95	21.2	399.6	0.04	-0.07
	3	215.5	71.13	20.9	402.0	-0.01	-0.42
	4	213.4	73.95	21.4	385.4	-0.08	-0.37

By Changing the extreme values, the outliers are generated. A typical scheme of changed extreme values by sample size are shown in Table 5. When lowest and highest values are changed, the effort was to assess whether the chosen method identify them as outliers.

Table 5: Scheme of Changed Extreme Values for one Typical Sample by Sample size

Sample Size	Status of Value	Original	Changed Extreme Values (Outliers)				
40	Min	21.5	19.5	17.5	15.5	13.5	11.5
	Max	73.5	75.5	77.5	79.5	81.5	83.5
60	Min	9.9	7.9	5.9	3.9	1.9	0
	Max	111.1	113.1	115.1	117.1	119.1	121.1
100	Min	26.9	24.9	22.9	20.9	18.9	16.9
	Max	150.3	152.3	154.3	156.3	158.3	160.3
200	Min	18.0	15.0	12.0	9.0	6.0	0.0
	Max	206.4	209.4	212.4	215.4	218.4	221.4
300	Min	35.0	28.0	21.0	14.0	7.0	0.0
	Max	453.9	460.9	467.9	474.9	481.9	490.9

Table 6 displays the original and the modified values, along with the corresponding Outlier identification method. Out of four changed lowest values namely 11.5, 13.1, 15.9 and 17.1, IQR-Old method and SD-Ranged method could identify only 11.5 as the outlier and missed the remaining three values in identification of them as Outliers. In contrast, IQR-Takiar method could identify three out of four values as Outliers. In case of highest values when changed for the four samples, IQR-Old method could identify only 108.6 as the Outlier while SD-Range-Takiar method could identify 108.6 and 95.8 as the Outliers. Among the three methods, IQR-Takiar method is the superior method to identify all the four values as Outliers. Thus, out of 8 Outliers to be identified, IQR-Old, SD-Range-Takiar and IQR-Takiar method could identify 2, 3 and 7 Outliers, respectively.

Table 6: Detection of Outliers by Samples and Methods- A typical Example

Status	Value	Sample 1	Sample 2	Sample 3	Sample 4
Original	Lowest Value	21.5	23.1	25.9	20.7
	Highest value	73.5	99.2	98.6	85.8
Changed	Lowest Value	11.5	13.1	15.9	10.7
	Highest value	83.5	109.2	108.6	95.8
Method	Value	Outlier Detected			
IQR-Old	Lowest value	11.5	-	-	-
	Highest value		-	108.6	-
IQR-Takiar	Lowest value	11.5	13.1	15.9	
	Highest value	83.5	109.2	108.6	95.8
SD-Range - Takiar	Lowest value	11.5	-	-	-
	Highest value	-	-	108.6	95.8

Detection of outliers by IQR-Old method by sample size and trial is shown in Table 7.

Table 7: Detection of Outliers by Sample size and Trial - IQR-Old Method

Trial	Sample size					Total
	40	60	100	200	300	
1	2	2	3	4	6	17
2	2	2	3	3	6	16
3	3	5	5	5	6	24
4	2	6	6	5	7	26
5	2	7	6	6	7	28
Pooled	11	22	23	23	32	111
% Outliers detected	27.5	55	57.5	57.5	80.0	55.5
Outliers Introduced	40	40	40	40	40	200

In five trials, out of 200 Outliers introduced, IQR-Old method could detect only 111(55.5%) Outliers. The percentage detection of Outliers varied from 27.5% in sample size of 40 to 80.0% in samples with size of 300.

Detection of Outliers by SD-Range-Takiar method by sample size and trial is shown in Table 8.

SD-Range Takiar method could detect 79 (39.5%) of the Outliers, correctly. The percentage detection of Outliers varied from 27.5% in sample size of 40 to 55.0% in samples size of 100.

Detection of outliers by IQR-Takiar method by sample size and trial is shown in Table 9. This method could detect 136 (68.0%) of the outliers, correctly. The percentage detection of Outliers varied from 55.0% to 97.5% in samples.

Table 8: Detection of Outliers by Sample size and Trial - SD-Range-Takiar Method

Trial	Sample size					Total
	40	60	100	200	300	
1	1	2	4	2	1	10
2	2	2	4	2	1	11
3	2	2	4	2	4	12
4	3	4	5	4	5	21
5	3	6	5	4	7	25
Pooled	11	16	22	14	18	79
% Outliers detected	27.5	40	55	35.0	45.0	39.5
Outliers Introduced	40	40	40	40	40	200

Discussion

The study attempted to evaluate three methods for detection of Outliers from the normal samples of sizes 40, 60, 100, 200 and 300.

Generally, for the normal samples, it is known that 99% of the observations lies in Mean ± 3SD.Thus, defining any observations outside the sample range as Outliers is not surprising and can be considered as the most appropriate and objective way of defining the Outliers.

Table 9: Detection of Outliers by Sample size and Trial - IQR-Takiar Method

Trial	Sample size					Sample size
	40	60	100	200	300	
1	2	7	5	4	1	19
2	3	8	6	3	3	23
3	5	8	6	4	6	29
4	5	8	7	5	6	31
5	7	8	7	6	6	34
Pooled	22	39	31	22	22	136
% Outliers detected	55.0	97.5	77.5	55.0	55.0	68.0
Outliers Introduced	40	40	40	40	40	200

One of the problem faced in identification of the Outliers is that different methods give different set of outliers even when same sample or samples are considered. When the number of Outliers picked up differ markedly by different methods, it become difficult for the researcher to decide which method is to be taken as the appropriate one. By defining the Outliers by an objective way, the present study has overcome the problem of subjectivity in the determination of the method to be the most appropriate. The method which picks up the larger number of Outliers outside the expected range can be taken as the most appropriate method.

In model building the presence of Outliers may pose a big problem and may reduce the utility of model building. Sometimes, deletion of some observations may enhance the R^2 value significantly. So, what we call as Outliers are actually not Outliers sometimes but they are either overrepresented or underrepresented values in the sample. Consider in a class of 30 students, a student having a height of 185 cm or a student of 120 kgs. Though for all practical purposes they are the part of our population but in derivation of mean height or weight of the class to consider their weight and height is not appropriate. Consideration of those two values may give inflated estimates of mean height and weight of the class.

In the present study, 200 outliers are introduced, spread over the four samples, five sample size and the five trials. IQR-Old method is extensively used to define the outliers particularly with the help of box plot by SPSS as well as Excel. This method has shown to pick up only 55.5% of the outliers out of 200 introduced. SD-Range method has shown to pick up only 39.5% of the Outliers (Fig. 1).

The inferior performance of the SD-Range-Takiar method is not surprising as in the presence of Outliers, the SD is inflated and the expected range based on the formula used would become larger enough to incorporate even the genuine outliers. IQR-Takiar method has shown to pick up 68% of the Outliers, a much higher percentage, as compared to remaining two methods.

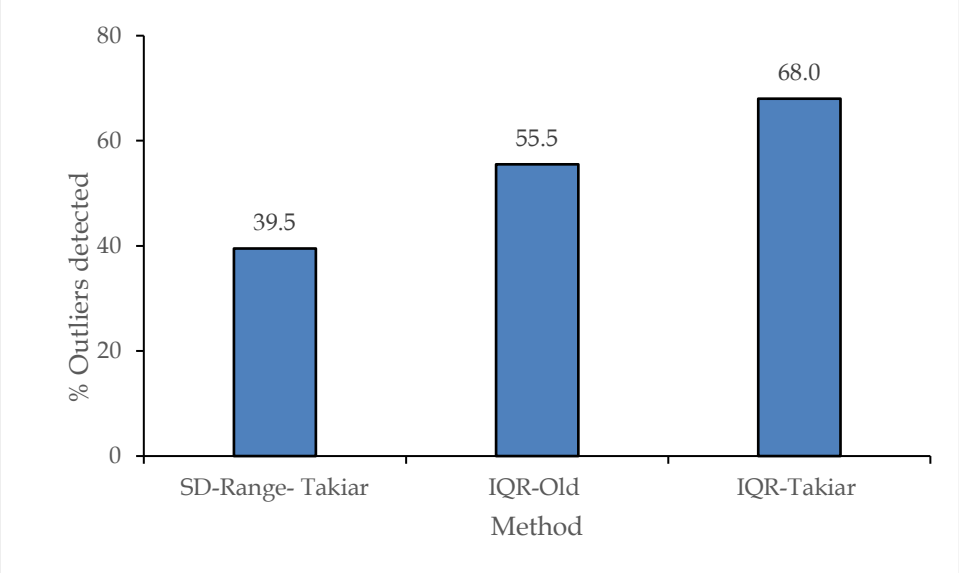


Fig. 1: % of Outliers detected by Different Methods

Thus, IQR-Takiar method is proved to be a superior method. Based on the study results, it is recommended to use IQR-Takiar method for detection of Outliers from large normal samples.

Summary of Observations

- In the present study, three method are used for the identification of Outliers in large normal samples.
- Four normal samples, each of size 40, 60,100,200 and 300 are utilized as a source of data.
- The three methods used are: IQR-Old method, IQR-Takiar method, SD-Range-Takiar method.
- The lower and higher fence values for three methods as defined are shown in the following table.

Method	Fence value	Formula for defining the Fence value
IQR-Old	LF	$Q_1 - 1.5*IQR$
	HF	$Q_3 + 1.5*IQR$
IQR-Takiar	LF	$Q_1 - IQR*[0.25*\ln(n)+0.20]$
	HF	$Q_3 + IQR*[0.25*\ln(n)+0.20]$
SD-Range -Takiar	LF	Mean - SD*(0.37*ln (n) + 0.86)
	HF	Mean + SD*(0.37*ln (n) + 0.86)

- For Validation purposes, all modified values defined outside the original range of a sample are treated as the Outliers.
- The Percentage detection rate of Outliers by the IQR-Old, the SD-Range and the IQR-Takiar method is observed to be 39.5%, 55.5% and 68.0%, respectively.
- Based on the study results, for large samples, the IQR-Takiar method is adjudged to be the superior method as compared to other two methods in detection of the Outliers.
- For small samples, for detection of outliers, it was already shown that IQR-Takiar method is a superior method.
- IQR-Takiar method is recommended for detection of Outliers in large samples.

References

[1]. Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research, 3*(1), 59–68. https://www.researchgate.net/publication/50946372_Outliers_detection_and_treatment_a_review

[2]. Hansen, J., Ahern, S., & Earnest, A. (2023). Evaluations of statistical methods for outlier detection when benchmarking in clinical registries: A systematic review. *BMJ Open, 13*, e069130. https://doi.org/10.1136/bmjopen-2022-069130

[3]. IBM Corp. (2015). *IBM SPSS Statistics for Windows* (Version 23) [Software].

[4]. Iacobucci, D., Román, S., Moon, S., & Rouziès, D. (2025). A tutorial on what to do with skewness, kurtosis, and outliers: New insights to help scholars conduct and defend their research. *Psychology & Marketing, 42*, 1398–1414. https://doi.org/10.1002/mar.22187

[5]. Microsoft. (2026). *Microsoft Excel* (Microsoft 365 MSO Version 2602, Build 16.0.19725.20126) [Software].

[6]. Onoz, B., & Oguz, B. (2003). Assessment of outliers in statistical data analysis. In N. B. Harmancioglu, S. D. Ozkul, O. Fistikoglu, & P. Geerders (Eds.), *Integrated technologies for environmental monitoring and information production* (NATO Science Series, Vol. 23). Springer. https://doi.org/10.1007/978-94-010-0231-8_13

[7]. Takiar, R. (2022). Sample variance—Is it really an unbiased estimate of the population variance? *Bulletin of Mathematics and Statistics Research, 10*(1), 21–30.

[8]. Takiar, R. (2023a). The relationship between the SD and the range and a method for the identification of outliers. *Bulletin of Mathematics and Statistics Research, 11*(4), 62–75.

[9]. Takiar, R. (2023b). A new method to identify outliers based on the interquartile method. *Bulletin of Mathematics and Statistics Research, 11*(4), 103–114.

[10]. Takiar, R. (2026). Comparison of the methods for detecting outliers. *Bulletin of Mathematics and Statistics Research, 14*(1), 25–34.

Biography

Dr. Ramnath Takiar

I am a Post graduate in Statistics from Osmania University, Hyderabad. I did my Ph.D. from Jai Narain Vyas University of Jodhpur, Jodhpur, while in service, as an external candidate. I worked as a research scientist (Statistician) for Indian Council of Medical Research from 1978 to 2013 and retired from the service as Scientist G (Director Grade Scientist). I am quite experienced in large scale data handling, data analysis and report writing. I have 81 research publications , with 1381 citations to my credit, published in national and International Journals related to various fields like Nutrition, Occupational Health, Fertility and Cancer epidemiology. During the tenure of my service, I attended three International conferences namely in Goiana (Brazil-2006), Sydney (Australia-2008) and Yokohoma (Japan-2010) and presented a paper in each. I also attended the Summer School related to Cancer Epidemiology (Modul I and Module II) conducted by International Agency for Research in Cancer (IARC), Lyon, France from 19th to 30th June 2007. After my retirement, I joined my son at Ulaanbaatar, Mongolia. I worked in Ulaanbaatar as a Professor and Consultant from 2013-2018 and was responsible for teaching and guiding the Ph.D. students. I also taugth Mathematics to undergraduates and Econometrics to MBA students. During my service there, I also acted as the Executive Editor for the in-house Journal "International Journal of Management". I am also acting as a reviewer for few International Journals. I am still active in research and have published 18 research papers in Statistical Methodologies during 2021-25.