Vol.11.Issue.4.2023 (Oct-Dec)

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

A NEW METHOD TO IDENTIFY THE OUTLIERS BASED ON THE INTER QUARTILE RANGE

RAMNATH TAKIAR

Apartment no. 11, Building 9, 3 rd Floor, Mangol Expoprt Town, 1st Khorro, Olympic Street Sukhbaatar District, Ulaanbaatar, Mongolia.

Email: ramnathtakiar@gmail.com, ramnath_takiar@yahoo.co.in

&

Scientist G – (Retired)

National Centre for Disease Informatics and Research (NCDIR), Indian Council of Medical Research (1978-2013) Bangalore – 562110, Karnataka, India.

DOI:10.33329/bomsr.11.4.103

Ramnath Takiar

ABSTRACT

The Outliers are such observations which behave differently in relation to rest of the observations. In case of single variable, the presence of Outliers may often result in skewed distribution. It may also inflate the mean and particularly the variance, very badly. Sometimes, in the presence of Outliers, the relationship studied between any two variables like income and expenditure may be altered significantly and may present a different form of relationship between them. Therefore, it is essential to know how to identify the Outliers in a data set and how to treat them for further analysis. In a recently published research paper, it was shown that the use of 1.5 as the multiplier of Interquartile Range (IQR), in identification of the Outliers is not so sensitive in picking up the expected Outliers. In the present paper, an attempt is made to find out a new method based on IQR for identification of the Outliers and comparisons are made with those Outliers which are identified by the SD-Takiar method.

From the data collected, based on the IQR-Takiar method, it was shown that the Lower Fence (LF) and Higher Fence (HF) can be defined as follows: LF = Q1 – IQR*[0.25*ln(n)+0.20] and HF= Q3+ IQR*[0.25*ln(n)+0.20]. While, the SD-Takiar method defines the LF and HF as follows: LF = Mean -

SD*(0.37*ln (n) + 0.86) and HF = Mean + SD*(0.37*ln (n) + 0.86). Thus, in the present study two new methods are given to identify the Outliers namely one is based on the SD of the sample another is based on the IQR of the sample. An agreement of 80% is found between the two methods, in identification of the Outliers.

Based on the finding that the relationship between the SD and the Range varies with the sample size, it is recommended to use varying cut-off levels for Z-test instead of using uniformly 1.96 as the cut-off point above the sample size of 30. The cut-off point may vary from 1.97 for the sample size of 20 to 2.31 for the sample size of 50 and 2.56 for the sample size of 100.

KEY WORDS: Outliers, Interquartile Range(IQR), IQR(Old) Method, IQR-Takiar method, SD-Takiar method, Varying cut-off points for Z-test.

INTRODCUTION

In research, data is collected as a routine or as a part of specific investigation. The data collected may be related to food intakes or anthropometric measurements or demographic variables like age, income, and expenditure. While analyzing the data so collected, sometimes, we may come across a very high or very low food intakes or high height measurements. Sometimes, the expenditure obtained is relatively very high as compared to the income of the family. Any unusual data, often draws the attention of the researchers and force them to verify whether the data so collected is correct and free from any technical or measurement errors? In literature, the extreme observations such as abnormally high or low observations in a data set, are termed as Outliers. Outliers are such observations which behave differently in relation to rest of the observations. In case of single variable, the presence of Outliers may often result in a skewed distribution and may affect the normality of the data. The presence of Outliers in a data set may also, inflate the mean, and particularly the variance, very badly. Sometimes, in the presence of Outliers, the relationship studied between any two variables like income and expenditure or age and mortality may be altered significantly and may present a different form of relationship between them. Therefore, it is essential to know how to identify the Outliers in a data set and how to treat them for further analysis.

There are articles which provide the details of few methods which can be used to identify the Outliers in a set of data. This includes Sorting of the data, Presentation of the data in Graphs, Calculation of Z scores of suspected data points, Calculation of Interquartile range, Use of Hypothesis testing (Graph Pad 2023, Bhandari P 2021, NIST 2012, Baraka S 2023, STAT200 2023, Frost J 2023). Sorting helps in visual identification of the data points that lies at the abnormal distance from rest of the values. When the data is presented in graph, it also helps in identification of the data points which are lying at the abnormal distance from the rest of the data points. Both the above methods suffer from the subjectivity in identification of the Outliers, involving the visual inspection and subjective decision of either excluding the Outliers or allowing those points to be retained in the data. The high or low Z scores of suspected data points, in relation to theoretically obtained critical Z scores, help us in deciding whether the suspected data points are genuinely Outliers or not? In Interquartile Range method, the distance of 1.5 times IQR from the first quartile (Q1) and third quartile (Q3) define the Lower fence and Upper fence values, respectively and any value lying outside these fence values are considered as the Outliers. In hypothesis testing one point at a time is considered for deciding whether it is an Outlier or not?

In a recently published research paper (Takiar R 2023), it was shown that the use of 1.5 as the multiplier of Interquartile Range (IQR), in identification of the Outliers, tend to miss at least 50% of

the Outliers as compared to the method developed by Takiar and named as SD-Takiar method. The Outliers missed by the IQR method are shown to be having Z values uniformly over 2.0. Thus, showing that the Interquartile method is not so sensitive in picking up the Outliers and needs change in the multiplier of 1.5 to some other suitable number. SD-Takiar method is relatively shown to be better in identification of the Outliers as compared to the existing Interquartile Range method. In the present study, an attempt is therefore made to develop a method based on the Inter quartile range to identify the Outliers in a data set and its comparisons are made with that of SD-Takiar method.

OBJECTIVES

1. To develop a new method based on the Inter Quartile Range to identify the Outliers.
2. For a selected set of data, identify the set of Outliers by newly developed method, based on the Inter Quartile Range and compare it with that obtained by the SD-Takiar method.

MATERIALS AND METHODS

DESCRIPTION OF THE NORMAL POPULATIONS

For the study purposes, it was decided to consider four types of Normal populations with predefined mean and the SD and are shown in Table 1.

Table 1: Description of Normal Populations with Specified Mean and SD

POPULATION	P1	P2	P3	P4
MEAN	60	80	100	120
SD	15	24	35	48

From each population, mentioned above, random samples with varying size like 15, 30, 50, 75, 100, 125, 150 and 200 are generated, and used for the study purposes. From each population and sample size, 15 samples are generated and denoted as S15, S30, S50, S75, S100, S125, S150, S200, respectively, resulting into a total of 60 samples for each Normal population. For more details about the selection of samples and populations, please refer to my previous study (Takiar R 2023).

DATA COLLECTED

From each sample, the following statistics were collected: Sample size (n), Mean, SD, Minimum value (MIN), Maximum (MAX), 1st Quartile (Q1), 3rd Quartile (Q3).

SD-TAKIAR METHOD FOR IDENTIFICATION OF THE OUTLIERS

In this method, based on the SD-Range relationship, the Lower Fence and Higher Fence values are defined and any values lying outside these Fence values are treated as the Outliers. It is to be noted that in case of SD calculation, the denominator used is ‘n’ and not ‘(n-1)’.

Lower fence value = LF = Mean - SD*{0.37*ln (n) + 0.86} Where n is the sample size

Higher fence value = HF = Mean + SD*{0.37*ln (n) + 0.86}

0LD INTER QUARTILE RANGE METHOD FOR IDENTIFICATION OF THE OUTLIERS

In this method, based on the Interquartile Range (IQR), the Lower Fence and Higher Fence values are defined and any values lying outside these Fence values are treated as the Outliers.

LF = Q1 – 1.5 * IQR; HF = Q3 + 1.5 * IQR Where IQR = (Q3 – Q1)

NEW INTERQUARTILE RANGE METHOD (IQR-Takiar) FOR IDENTIFICATION OF THE OUTLIERS

In this method, based on Inter quartile range, the Lower Fence (LF) and Higher Fence values are going to be redefined and any values lying outside these Fence values will be treated as the Outliers. The value of M will be determined based on the data selected.

Lower Fence value = LF = Q1 – M * IQR Higher Fence value = HF = Q3 + M * IQR

DETERMINATION OF M BASED ON Q1, Q3, MINIMUM AND MAXIMUM VALUE

Ideally, if we can estimate the Range of the data and can find some observations lying outside the estimated range then those points can be treated as the Outliers. For each sample size, 60 pairs of values of [IQR and (Q1-MIN)] and [IQR and (MAX-Q3)] are obtained and Linear regression models are generated. This allows us to estimate the distance between (Q1, MIN) and (MAX, Q3) in terms of IQR. Regression models generated, give us the value of slope (M) assuming the linear regression model represented as

IQR = M1*(Q1-MIN) + C1 IQR = M2*(MAX–Q3) + C2

The values of M1, M2 are likely to vary with the varying sample size. Keeping in view the objective of identifying the Outliers, the value of M1 and M2 are considered. Plotting the values of M1 and M2, according to different sample sizes, allow us to explore the relationship of M1 and M2 with that of the sample size.

INTERNAL AND EXTERNAL VALIDITY OF THE MODEL DEVELOPED

Validity of the model: It refers to the ability of the model (R^2) to assess correctly the relationship between the IQR and the distance between (Q1-MIN) on one hand and IQR and the distance between (MAX-Q3) on the other hand.

Internal Validity: To assess the internal validity, the value of "R" that is the correlation coefficient is obtained between IQR and the distance between (Q1-MIN) on one hand and IQR and the distance between (MAX-Q3) on the other hand. The higher value of R^2 (say more than 0.9) allow us to place the higher confidence in the model fitted and consider it to be more valid.

External Validity: It is desirable that the model developed is found to be applicable to other sets of data. The model developed is applied to the external data, for identification of the Outliers, to check the external validity of the model. A model with the proven external validity can safely be applied to any other sets of data.

ANALYSIS OF THE DATA

Keeping in view the objectives of the study, based on 60 samples drawn from the four Normal Populations, the regression models are generated involving IQR with (Q1-MIN) on one hand IQR with (MAX-Q3) on the other hand. To test the external validity of the models developed, 40 samples, 10 samples each, collected from the population of P5, P6, P7 and P8, are subjected to identifications of the Outliers by IQR-Takiar and IQR(Old) method and comparisons are made with that of SD-Takiar method to test their validity.

RESULTS

The results of the Regression analysis attempted for two pairs of data namely IQR and (Q1-MIN) for the varying sample size is shown in Table 2. Besides, giving the slope value (M1), the values of the Lower Confidence Limit (LCL) and Upper Confidence Limit (UCL) are also provided.

Table 2: The Regression Coefficient (M1) between IQR and (Q1-MIN) According to Varying Sample Size

SAMPLE SIZE	No. of Samples	M1	LCL	UCL	R	R^2
15	60	0.75	0.63	0.88	0.85	0.72
30	60	1.06	0.93	1.18	0.91	0.83
50	60	1.13	1.03	1.23	0.95	0.90
75	60	1.22	1.12	1.32	0.95	0.90
100	60	1.28	1.19	1.38	0.97	0.94
125	60	1.48	1.37	1.59	0.96	0.92
150	60	1.38	1.29	1.46	0.97	0.94
200	60	1.58	1.50	1.66	0.98	0.96

It is clear from the table that M1 values varies with the sample size suggesting that the relationship between (Q1-MIN) and IQR is a function of the sample size. Uniformly higher values of R^2 for varying sample size suggests that the model fitted is quite good.

An attempt is made to understand the extent of linear relationship between M1 and the sample size. The model fitted to the values of M1, is shown in Fig. 1. The higher value of R^2 (0.9491) suggests that the model fitted is quite good.

The results of the Regression analysis, attempted for two pairs of data namely IQR and (MAX-Q3) for the varying sample size is shown in Table 3. Besides, giving the slope value (M2), the values of the Lower Confidence Limit (LCL) and Upper Confidence Limit (UCL) are also provided. M2 values, like M1 values, are also seen to be varying with the sample size suggesting that the relationship between (MAX -Q3) and IQR is a function of the sample size. A high value of R^2 suggests that the model fitted is quite good.

Fig. 1: The Relationship between (Q1-MIN) and Multiplier (M1) of IQR

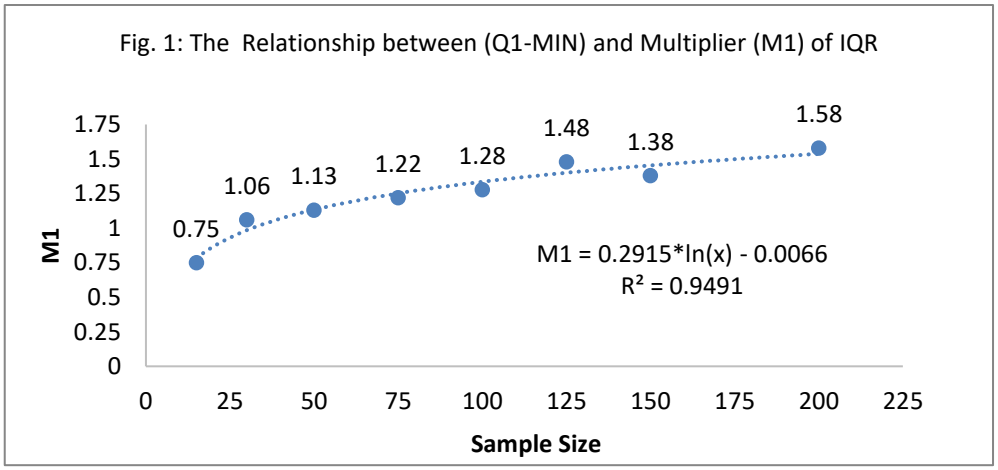

Table 3: The Regression Coefficient (M2) between IQR and (MAX-Q3) According to Varying Sample Size

SAMPLE SIZE	No. of Samples	M2	LCL	UCL	R	R^2
15	60	0.89	0.75	1.03	0.86	0.74
30	60	1.03	0.92	1.14	0.92	0.85
50	60	1.21	1.12	1.30	0.96	0.92
75	60	1.29	1.18	1.40	0.95	0.90
100	60	1.37	1.29	1.46	0.97	0.94
125	60	1.47	1.36	1.58	0.96	0.92
150	60	1.50	1.41	1.59	0.98	0.96
200	60	1.48	1.41	1.56	0.98	0.96

An extent of linear relationship between M2 and the sample size is attempted and shown in Fig. 2. The higher value of R^2 (0.9635) suggests that the model fitted is quite good.

Based on the models shown in Fig. 1 and Fig. 2, the actual values, and the model values of M1 and M2 are shown in Table 4. As shown in the table, the model fitted values are quite close to those obtained by varying sample size. It is decided to choose that model which gives the expected values of (M1, M2) as the maximum corresponding to the selected sample sizes. Accordingly, the model shown in Fig. 2 for identification of the Outliers.

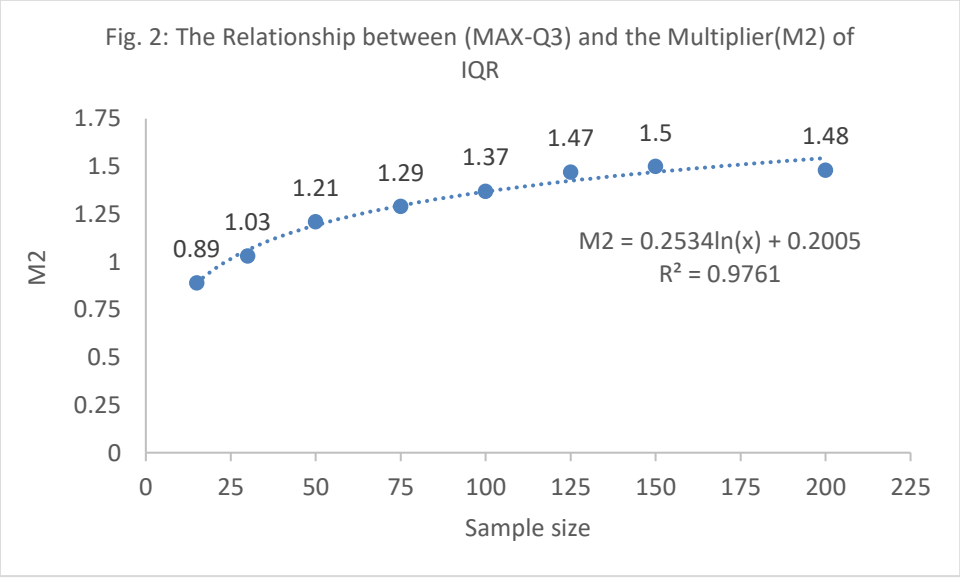

Table 4: The M1 and M2 Values and Their Model Fitted values

N	M1	Expected	M2	Expected
15	0.75	0.79	0.89	0.89
30	1.06	0.99	1.03	1.06
50	1.13	1.14	1.21	1.19
75	1.22	1.26	1.29	1.29
100	1.28	1.34	1.37	1.37
125	1.48	1.41	1.47	1.42
150	1.38	1.46	1.5	1.47
200	1.58	1.54	1.48	1.54

The 20 random Normal samples are generated and considered for the testing the External Validity of the IQR-Takiar method in identification of the Outliers and compared with those obtained by the SD-Takiar method (Takiar R 2023).

The 10 random samples (S1-S10) with the identified Outliers are shown in Table 4A and the remaining 10 samples (S11-S20) with correspondingly identified Outliers are shown in Table 4B. The list of Outliers identified by both the methods are shown in the Table 5.

Table 4A: Sets of Data with Identified Outliers by SD based and IQR based Method

No.	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1	**30.40**	**30.96**	**25.12**	**40.34**	**23.03**	**50.26**	35.80	**39.65**	75.33	**31.45**
2	49.78	39.47	45.64	48.63	43.61	57.58	38.90	42.13	76.31	39.53
3	51.30	42.91	55.68	54.51	48.67	60.07	48.96	52.79	78.32	44.95
4	51.55	44.19	55.87	56.38	49.32	68.59	49.25	68.73	83.43	55.97
5	55.05	52.79	59.80	56.57	50.04	68.61	51.60	71.74	84.53	59.05
6	58.74	56.51	60.99	57.52	55.39	68.80	51.99	72.59	87.11	65.52
7	59.51	58.45	68.59	58.29	60.08	68.95	54.68	75.37	88.43	66.04
8	61.09	58.86	69.66	66.91	60.88	69.13	56.15	78.25	90.80	88.99
9	62.56	59.60	73.45	69.16	62.44	69.53	58.33	89.74	91.59	89.87
10	64.30	63.30	74.21	72.90	63.29	75.73	69.46	90.79	93.02	90.47
11	64.50	66.65	75.04	73.13	64.75	78.73	70.21	95.34	103.45	92.69
12	66.00	69.27	76.22	75.68	64.82	78.91	74.34	95.97	106.49	96.60
13	69.05	69.71	76.81	79.57	69.26	79.31	78.54	97.14	108.06	97.36
14	73.41	70.02	78.37	84.42	69.26	81.64	87.10	98.90	111.15	101.56
15	75.23	70.07	80.74	84.44	72.50	87.45	90.03	100.85	128.66	104.29
16	77.34	72.32	84.35	87.17	73.80	87.87	90.49	110.71	135.32	109.26
17	84.46	73.38	85.74	87.61	81.10	89.66	96.64	111.76	138.78	110.28
18	86.56	86.48	92.34	88.45	82.55	90.51	97.47	116.49	139.37	113.68
19	88.53	86.82	92.80	91.72	86.01	92.34	111.71	117.20	152.66	115.73
20	**96.60**	**90.03**	**108.32**	93.55	**112.26**	95.27	**120.97**	**167.18**	**177.97**	129.62

The 10 random samples (S11-S20) with identified Outliers are shown in Table 4B. The SD-Takiar method identifies 29 Outliers. Out of which 23 (79.3%) are also identified by IQR-Takiar method. The SD-Takiar method, thus identified 6 extra Outliers as compared to IQR-Takiar method. The IQR-Takiar method picked up 3 out of 4 extra Outliers with low Z values, below 1.96.

Table 4B: Sets of Data with Identified Outliers by SD based and IQR based Method

No.	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20
1	**31.77**	**46.60**	**36.19**	**39.02**	**30.94**	88.51	**17.64**	49.84	**36.10**	**34.71**
2	46.82	**52.09**	62.15	51.48	56.78	96.70	54.72	72.83	51.30	62.64
3	57.06	75.57	77.66	52.54	60.44	97.59	71.03	73.72	80.11	71.24
4	62.13	77.16	77.86	58.81	73.21	100.76	76.46	81.40	83.55	75.03
5	68.62	83.13	78.30	67.87	75.05	110.48	79.98	82.67	84.05	79.53
6	79.93	87.62	85.25	69.72	79.60	113.64	80.47	87.82	89.57	102.85
7	86.56	95.84	92.16	71.42	88.52	115.45	86.06	94.41	93.48	111.60
8	93.44	97.49	93.22	74.53	91.86	120.75	88.08	102.61	93.87	118.01
9	103.22	98.84	93.62	75.22	101.38	125.88	100.30	106.88	102.21	119.58
10	104.01	100.46	93.91	75.87	101.84	134.90	102.28	108.70	102.29	120.75
11	104.62	101.54	96.96	79.04	105.73	140.77	106.00	110.97	112.78	123.85
12	111.29	102.21	97.15	85.99	108.12	152.06	111.28	116.38	119.12	125.14
13	112.84	109.84	98.38	88.97	110.13	157.38	115.69	122.73	129.54	130.21
14	113.01	113.07	99.59	89.79	113.63	163.43	116.99	128.47	130.93	140.16
15	119.63	113.82	112.62	99.44	119.20	169.12	124.25	129.37	132.53	145.08
16	124.78	116.68	114.95	103.42	120.88	175.34	132.68	139.06	140.49	153.92
17	126.81	117.64	121.58	106.09	124.66	181.39	139.24	150.66	152.77	157.09
18	127.81	119.06	128.66	108.23	127.38	183.75	148.71	177.20	157.81	170.04
19	128.96	127.41	134.97	109.18	134.22	192.35	163.95	**182.56**	160.08	184.81
20	**177.99**	**150.94**	**168.91**	117.70	**165.80**	**234.29**	**181.49**	**222.70**	164.59	**211.27**

Table 5: The List of Outliers Identified by the IQR-Takiar and the SD-Takiar Method

COMMON OUTLIERS				OTHER OUTLIERS	
OUTLIER	Z VALUE	OUTLIER	Z VALUE	OUTLIER	Z VALUE
30.4	-2.34	150.94	2.15	40.34*	-2.01
96.6	1.98	36.19	-2.27	120.97*	2.09
30.96	-2.08	168.91	2.58	31.45*	-1.97
25.12	-2.62	30.94	-2.27	31.77*	-2.01
108.32	2.03	165.80	2.20	39.02*	-1.99
23.03	-2.31	234.29	2.41	34.71*	-2.06
112.26	2.64	17.64	-2.33	90.03**	1.74
50.26	-2.12	181.49	2.05	39.65**	-1.75
167.18	2.72	222.70	2.58	52.09**	-1.97
177.97	2.51	36.10	-2.15	182.56**	1.60
177.99	2.36	211.27	2.12		
46.60	-2.20				

*- Only by the SD-TAKIAR Method; ** - Only by the IQR-TAKIAR Method.

The IQR-Old method identifies only 5 outliers as against 29 Outliers identified by the SD-Takiar method and 27 Outliers identified by the IQR-Takiar method. This clearly shows that the method is not sensitive in picking up the Outliers. The list of Outliers picked up by the IQR(Old) method is not shown here.

DISCUSSION

For the study purposes, sixty samples are generated for the varying sample size like 15, 30, 50, 75, 100, 125, 150 and 200. For each sample size, 60 pairs of values of [IQR and (Q1-MIN)] and [IQR and (MAX-Q3)] are obtained and Linear regression models are generated. This allows us to estimate the distance between (Q1 and MIN) and (MAX and Q3) in terms of IQR. Thus, the regression analysis give us two estimates of distance from IQR, one that is emerging from (Q1 and MIN) and another from (MAX and Q3). Both the estimates are important for identifying the Lower Fence Value (LF) and Higher Fence Value (HF), respectively, so that any points lying outside these intervals are considered as Outliers. Ideally speaking, if the samples are symmetric towards the mean, the multiplier emerging from LF and HF should be comparable.

It is common knowledge that though the samples are drawn from a normally distributed population, they may not be symmetrical towards the mean. To have a unique multiplier of IQR to measure the distance of MIN from Q1 (M1) and MAX from Q3 (M2), the model which gives the maximum of the expected values of (M1, M2) is considered. Accordingly, the model corresponding to M2, shown in Table 3 is selected and used for identification of the Outliers. The model chosen is **M2 = 0.25*ln(n) + 0.20**. Depending on the sample size, the multiplier to IQR can be calculated using the above model equation. For easy reference, few Multiplier values of IQR are calculated for the selected sample sizes and shown in Table 6. For other sample sizes, below100, the Multiplier to IQR can be interpolated.

Table 6: The Multiplier of IQR Estimated from the Regression Equation for the Identification of the Outliers by IQR-Takiar Method

SAMPLE SIZE	10	15	20	30	40	50	60	70	80	100
MULTIPLIER	0.78	0.88	0.95	1.05	1.12	1.18	1.22	1.26	1.30	1.35

In the IQR(Old) method, based on Interquartile Range (IQR), the Lower Fence and Higher Fence values are defined as **LF = Q1 – 1.5 * IQR; HF = Q3 + 1.5 * IQR**. The fact that the Multiplier of IQR is a function of the sample size, to use uniformly 1.5 as the multiplier of IQR is not appropriate. Further, the use of 1.5 as the multiplier of IQR results in longer interval of (LF-HF) than the expected, resulting in picking up the less percentage of genuine Outliers.

In my earlier study, it was shown that IQR(Old) method tend to pick up only 50% of that picked up by the SD-Takiar Method. Based on the findings there, it was suggested that there is a definite need of identifying another multiplier of IQR to identify correctly the Outliers. The current study has shown that the multiplier cannot be a constant value like 1.5 for IQR and is observed to be a function of the sample size. As evident from the Table 6, the Multiplier remained below 1.5 even for the sample size of 100, suggesting that the old method of IQR is not appropriate and need to be replaced with the value suggested by the IQR-Takiar method. Accordingly, for the identification of the Outliers, the LF and HF are defined in terms of IQR as follows:

LF = Q1- IQR*[0.25*ln (n) + 0.20] and HF = Q3 + IQR*[0.25*ln (n) + 0.20]

In my recent study (Takiar R 2023), it was shown that the SD and the Range are related to each other and the relationship changes with the change in the sample size. The relationship between them can be characterized by the equation; Range=SD*(0.73*ln(n) + 1.72). Assuming that the range is equally distributed around the mean, the Lower fence (LF) and Higher Fence (HF) values, to identify the Outliers are defined as follows:

LF = Mean - SD*(0.37*ln (n) + 0.86) and HF = Mean + SD*(0.37*ln (n) + 0.86)

For easy reference, few values are calculated according SD-Takiar method for selected sample sizes and shown in Table 7. For other sample sizes below100, the multiplier of SD can be interpolated.

Table 7: The Multiplier of SD Estimated from the Regression Equation for the Identification of the Outliers by the SD- Takiar Method

SAMPLE SIZE	10	15	20	30	40	50	60	70	80	100
MULTIPLIER	1.71	1.86	1.97	2.12	2.22	2.31	2.37	2.43	2.48	2.56

The fact that the Multiplier to SD changes even after the sample size of 30, suggests that for Z-test to use uniformly 1.96 as the cut-off point for testing the significance for single mean or between two means is not appropriate and can be considered according to the sample size chosen. If the theory is correct, the multiplier should not be varying much with the sample size but it is varying and varying from 1.97 for the sample size 20 to 2.31 for the sample size of 50 and 2.56 for the sample size of 100. Based on this finding, it is appropriate to think that a uniform cut off level of 1.96 should be abandon for normal samples and a cut-off level based on the sample size to be used, hereafter.

The 20 random Normal samples are generated and considered for the identification of the Outliers by the IQR-Takiar Method and by the SD-Takiar Method. The SD-Takiar method identifies 29 Outliers. Out of which 23 (79.3%) are also identified by the IQR-Takiar method. The SD-Takiar method, thus identified 6 extra Outliers as compared to IQR-Takiar method but all those values exhibited higher Z values than 1.97, thereby justifying their inclusion in the list of Outliers. There appears to be about 80% agreement between both the methods in picking up the Outliers.

CONCLUSIONS

- The current study has highlighted two methods, developed, and found to be suitable with proven external validity, for identification of the Outliers.
- IQR-Takiar method, a new method developed in the current study, redefine the LF and HF values for identification of the Outliers as follows:

LF = Q1- IQR*{0.25*ln (n) + 0.20} and HF = Q3 + IQR*{0.25*ln (n) + 0.20}

- For quick reference of the Multiplier of IQR, for the sample size between 10 and 100 can be referred from the Table 6.
- For the identification of the Outliers based on the SD of the sample, the SD-Takiar method define the LF and HF values as follows:

LF = Mean - SD*{0.37*ln (n) + 0.86} and HF = Mean + SD*{0.37*ln (n) + 0.86}

- For quick reference of the Multiplier of SD, for the sample size between 10 and 100 can be referred from the Table 7.
- Based on the 20 Normal samples generated, the SD-Takiar method identifies 29 Outliers.
- The IQR-Takiar method identifies 27 Outliers.
- There appears to be 80% agreement in both the methods for identification of the Outliers.
- The IQR-Old method identifies only 5 outliers as against 29 Outliers identified by the SD-Takiar method and 27 Outliers identified by the IQR-Takiar method.
- The IQR-Old method, appears to be less sensitive in identifications of the real Outliers and thus need to be replaced with the newly developed IQR-Takiar method.

RECOMMENDATIONS

- Based on the IQR-Takiar method developed in the current study, the LF and HF values, for identification of the Outliers, are defined as follows:

LF = Q1- IQR*(0.25*ln (n) + 0.20) and HF = Q3 + IQR*(0.25*ln (n) + 0.20)

- Based on the SD-Takiar method, define the LF and HF values as follows:

LF = Mean - SD*(0.37*ln (n) + 0.86) and HF = Mean + SD*(0.37*ln (n) + 0.86)

- Based on the findings of the current study, It is recommended to use the varying cut-off points, based on the sample size, for Z-test, instead of using uniformly 1.96 as the cut-off point.
- The cut-off point may vary from 1.97 for the sample size of 20 to 2.31 for the sample size of 50 and 2.56 for the sample size of 100.

REFERERENCE

[1]. Baraka S 2023: What are Outliers in Statistics? Plus 5 ways to Find them https://www.indeed.com/career-advice/career-development/outliers-statistics

[2]. Bhandari P 2021: How to Find Outliers, 4 ways with Examples and Explanation https://www.scribbr.com/statistics/outliers/

[3]. Frost J 2023: 5 Ways to Find Outliers in Your Data, Statistics by Jim https://www.statisticsbyjim.com/basics/outliers

[4]. GraphPad 2023: Outlier Calculator

https://www.graphpad.com › quickcalcs › Grubbs1

[5]. NIST/SEMATECH e-Handbook of Statistical Methods 2012: What are outliers in the data

https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

[6]. [STAT200/Elementary Statistics 2023: Identifying Outliers: IQR Method, Penn State Eberly college of Science https://online.stat.psu.edu/stat200/lesson/3/3.2

[7]. Takiar R 2023: The Relationship between the SD and the Range and a method for the Identification of the Outliers; Bulletin of Mathematics and Statistics, Vol. 11(4), 62-75; KY Publications, India. DOI:10.33329/bomsr.11.4.62

Biography of corresponding author: 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 65 research publications 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 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 still active in research and have published 12 research papers during 2021-23.