Vol.11.Issue.4.2023 (Oct-Dec)
©KY PUBLICATIONS

http://www.bomsr.com

Email:editorbomsr@gmail.com

RESEARCH ARTICLE

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

The Relationship between the SD and the Range and a method for the Identification of the Outliers

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.11.4.62

Ramnath Takiar

ABSTRACT

The Standard Deviation and the Range are meant to study the variation in the data. Therefore, it is logical to expect that they are related to each other. It is further known that Inter-quartile range is used to identify the Outliers. Like, Inter-quartile range, it is also possible to develop an equation in terms of the Standard deviation, in identification of the Outliers. In the present paper, an attempt is made to develop a model to study the relationship between the SD and the Range. Further, an attempt is also made to develop a method based on the Mean and SD to identify the Outliers in a series of data?

For the study purposes, it was decided to consider four types of Normal populations with predefined mean and the SD. For each population, 15 samples, of size 15, 30, 50, 75, 100, 125, 150 and 200 are generated. Thus, for each Normal Population, 60 samples are obtained. For each sample, Mean, SD and the Range are calculated. The data so collected is analyzed keeping in view the objectives of the study. The analysis revealed that there exists a relationship between the SD and the Range of the data and which 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). The establishment of the External Validity of the model proves that it is suitable for application to any external data. The method developed has also shown to be better in picking up of the Outliers, 50% more, as compared to the that picked up by the Inter Quartile Range (IQR) method. The IQR method assign, on an average, 35% higher fence size, as compared to Mean and SD Method. There is a need to validate the 1.5 as the Multiplier of IQR, to achieve the lower and higher fence values in identification of the Outliers.

Keywords: Range, SD, Relationship, Outliers, Inter Quartile Range, Mean-SD Method.

INTRODUCTION

A set of the data can be characterized largely by a measure of central tendency and a measure of dispersion. A measure of central tendency is used to form an idea about the central part of the distribution while a measure of dispersion is used to form an idea about the scatteredness of the data. The most popular measures of central tendency being the Mean. Among the measures of dispersion, the Standard Deviation, Interquartile Range, and the Range are in common use. A low or small Standard Deviation indicates that the most of the observations in the data set are lying close to the mean while a high or large Standard Deviation indicates that the observations are quite scattered and are lying far from the mean. In case of a scientific data or a numeric data, the Standard Deviation is often used to get an idea about the spread of the observations in the data.

The Range depicts the difference between two extreme observations. It has found a quite many applications like in stock market, money rates and rates of exchange. In industries, the Range is used to define the quality control measures of the manufactured products by defining the control charts. In Medical reports, the range is often used to define the normal levels of the clinical parameters like blood count, Hb level and Cholesterol level. In Weather forecast, for a given day, a range of the temperature is reported to provide an idea about the temperature pattern of the area. For an area, the rainfall is also reported in terms of range.

The Standard Deviation and the Range are meant to study the variation in the data. Therefore, it is logical to expect that they are related to each other. There are few articles to show that the Range and the Standard Deviation are related, describing vaguely by the thumb rule equation (Taylor C 2019, Jim 2023, Schenkelberg F 2023) that Standard Deviation = $\frac{(Max-Min)}{4}$. The logical basis for this comes from the fact that for any normal sample, we can expect that 95% of the observations are lying between Mean $\pm$ 2SD, thus the Range can be approximated by 4SD or the Range divided by 4 can be taken as a rough approximation of the SD.

In the presence of outliers or extreme values in the data, the mean and median do not remain the same and the distribution of the data appear to be skewed and far from being the normal. In such a situation, the use of median is preferred over the mean. Accordingly, to form an idea about the variability in the skewed data, the use of the interquartile range is suggested. Unlike, the SD and the Range, the Inter quartile range is not affected by the extreme values or the Outliers.

It is further known that Inter-quartile range is used to identify the Outliers (Bhandari P 2021, Frost J 2023, Baraka S 2023, STAT200 2023). Like, Inter-quartile range, it is also possible to develop an equation in terms of the Standard deviation, in identification of the Outliers, for a given set of the data. The present paper is an attempt to view the possible relationship in above mentioned variables with the objective of answering the following questions:

- What is the relationship between the Standard Deviation and the Range?
- Is the relationship between the SD and the Range, changes with the size of the sample?
- How can the Standard Deviation be used to identify the Outliers, in a set of data?
- How do the Identification of Outliers compare between the IQR Method and the Mean-SD Method?

MATERIAL 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 the Normal Populations with the Specified Mean and the SD

POPULATION	P60	P80	P100	P120
MEAN	60	80	100	120
CV	0.25	0.3	0.35	0.4
SD	15	24	35	48

GENERATION OF RANDOM SAMPLES

For each population of P60, P80, P100 and P120, seven types of samples with the size of 15, 30, 50, 75, 100, 125, 150 and 200 are generated, randomly, using the Function available at StatPlus 7.6.5. To make use of the Function, one must proceed with Clicking the DATA → Random Number Generation → Normal →Number of New Variables → Random Number Count → Mean → Standard Deviation. Based on the input, the defined number of random samples of given size with the defined mean and SD, are generated. The number of samples selected by the type of population is shown in Table 2.

Table 2: Selection of Number of Samples by the Normal Populations

POPU - LATION	S15	S30	S50	S75	S100	S125	S150	S200
P60	15	15	15	15	15	15	15	15
P80	15	15	15	15	15	15	15	15
P100	15	15	15	15	15	15	15	15
P120	15	15	15	15	15	15	15	15
TOTAL	60	60	60	60	60	60	60	60

For each population, 15 samples, of size 15, 30, 50, 75, 100, 125, 150 and 200 are generated and denoted by S15, S30, S50, S75, S100, S125, S150, S200, respectively. Thus, for each Normal Population, 60 samples are obtained, as shown in the Table 2.

DATA GENERATED

For each sample, Mean, SD, the Minimum value (MIN) and the Maximum (MAX) value are noted. With the help of MIN and MAX values, the Range is calculated using the formula:

Range = MAX – MIN.

To test this hypothesis that there exists a relationship between the Range and the SD, the correlation and the regression equation are obtained from the pairs of SD and the Range data, generated from the 60 samples of each sample size. The scatter plots are also obtained and the respective Regression equation is displayed on the chart with R^2 value where R represents the correlation. A high R^2 value is assumed to support the hypothesis stated and the Range is estimated using the SD value.

DEFINITION OF OUTLIERS

In a numercal data, Outliers are such values which do not fit into the general pattern of the data. The Outliers may correspond to extremely low or extremely high values in the series of a data. The presence of Outliers in a set of data, may affect badly the shape or pattern of the data. In literature, the method of Interquartile Range is used to identify the Outliers (Bhandari P 2021, Frost J 2023, Baraka S 2023, NIST/SEMATECH 2012).

INTERQUARTILE METHOD TO IDENTIFY THE OUTLIERS

This method uses a range of values, defined with the help of first and third quartiles, to identify the Outliers. The formula make use to two values termed as Low fence and High fence values, beyond which any value lying is considered as an Outlier.

Inter Quartile Range = IQR = $(Q_3 – Q_1)$

Where Q_3 is the 3rd Quartile and Q_1 is the 1st Quartile

Low Fence value = LF = $Q_1 – 1.5 * \text{IQR}$

High Fence value = HF = $Q_3 + 1.5 * \text{IQR}$

DEVELOPMENT OF MEAN-SD METHOD FOR FINDING THE OUTLIERS

For any Normal sample, it is expected that the Mean ± 2SD will account for the 95% of the data. Similarly, Mean ± 3SD will account for the 99.7% of the data. For large samples, more than 30, it is believed to hold good. However, in case of small samples, say, with the sample size of 15, 95% will corresponds to accounting of 14.25 or 14 number of the observations. This amount to saying that for small samples, it is enough, if we consider all the observations lying outside Mean ± 2SD as the Outliers. However, with rise in the sample size, to use Mean ± 2SD as the criterion to define Outliers is subject to verification.

The SD and the Range, both being the measures of variation in the data, it is expected that they are related to each other. Therefore, if an equation can be developed based on the empirical relationship of the SD and the Range, it is possible for us to define a criterion for identification of the Outliers. Model Equation, so developed will be subjected to the internal validity as well as the external validity.

INTERNAL AND EXTERNAL VALIDITY OF THE MODEL DEVLOPED TO ASSESS THE RELATIONSHIP BETWEEN THE RANGE AND THE SD

- **Validity of the model:** It refers to the accuracy of the model in the assessment of the relationship between the SD and the Range. By the accuracy of the model, we refer to the ability of the model to assess correctly the relationship between the SD and the Range. Further, the presence of the confounding factors, if any, affecting the outcome of the model, are also assessed carefully.

- **Internal Validity**: The Internal Validity examines the level of accuracy that you can attach to the model developed to assess the relationship between the SD and the Range. In this connection, besides giving the linear equation depicting the relationship between the Range and the SD, the R^2 value is also provided. The higher the value of R^2, the higher the confidence one can place in the model fitted and consider it to be more valid.

- **External Validity:** It is not essential that a model developed based on the certain data collected, should also be applicable with the equal confidence to other sets of data. However, if the model is found to be applicable to other sets of data with equal confidence, then it proves its external validity. A model with the proven External validity, can safely be used to similar set of data. For the study purposes, four sets of Normal samples are generated with the specified Mean and the SD and subjected to testing the External Validity of the Model developed and shown in Table 3.

IDENTIFICATION OF OUTLIERS AND METHODS

Identification of the Outliers using IQR method is quite popular and can be found easily on the web For the selected sets of data (Table 3), Outliers will be identified by both the methods of IQR and SD-Range Method and the comparison will be made. A good percentage of the agreement in identification of the Outliers between both the methods, would serve as the basis of validation of SD based method for Identification of the Outliers.

Table 3: Four sets of Sample data generated with the Specified Mean and the SD for testing the External Validity of the Model developed

MEAN	CV	SD	S10	S20	S40	S80
70	0.25	17.50	10	10	10	10
90	0.30	27.00	10	10	10	10
110	0.35	38.50	10	10	10	10
130	0.40	52.00	10	10	10	10
Total			40	40	40	40

It is good to examine some other studies (Bhandari P 2023) for more interesting details about the Internal and the External Validity.

RESULTS

The relationship between the Range and the SD, based on the 60 pairs of the data, derived from the Normal samples of size 15, is shown in Fig. 1. The fact that $R^2 = 0.9447$, indicates that the Range can be estimated easily with the help of the regression equation of the Range and the SD, shown in the Fig. 1. The slope of 3.65, here, indicates that the Range is approximately 3.65 times of SD and that almost all the data is accounted, approximately, by the Mean ± 1.83*SD.

The relationship between the Range and the SD, based on the 60 pairs of data, derived from the Normal samples of size 30, shown in the Fig. 2. The $R^2 = 0.9147$ value indicates that the Range can be estimated easily with the help of the regression equation, shown in the Fig. 2. The slope of 4.10 suggests that almost all the data can be accounted by Mean ± 2.05*SD when the sample size is 30.

The relationship between the Range and the SD, based on the 60 pairs of data, derived from the 60 Normal samples of size 50, is shown in the Fig. 3. The value of $R^2 = 0.9345$, indicates that the Range can be estimated easily with the regression equation of the Range and the SD, shown in the Fig. 3. The slope of 4.54 obtained, suggests that it is possible to account for almost all observations by Mean ± 2.27*SD when the sample size 50.

Fig. 1: The Relationship Between the Range and the SD When the Sample size is 15 (n =60)

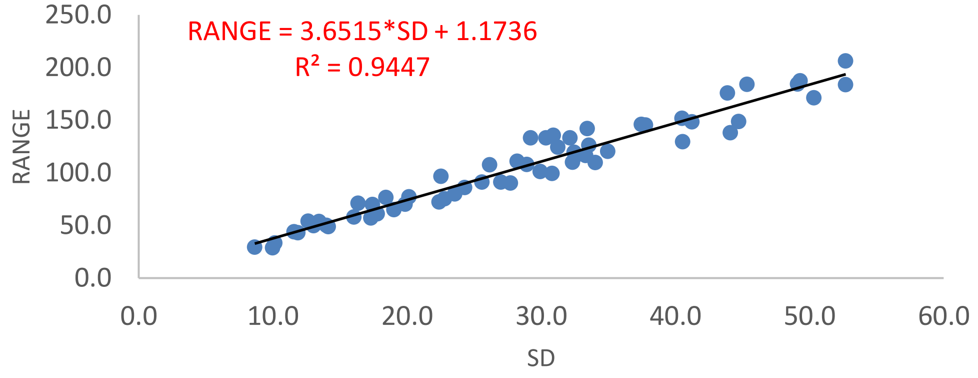
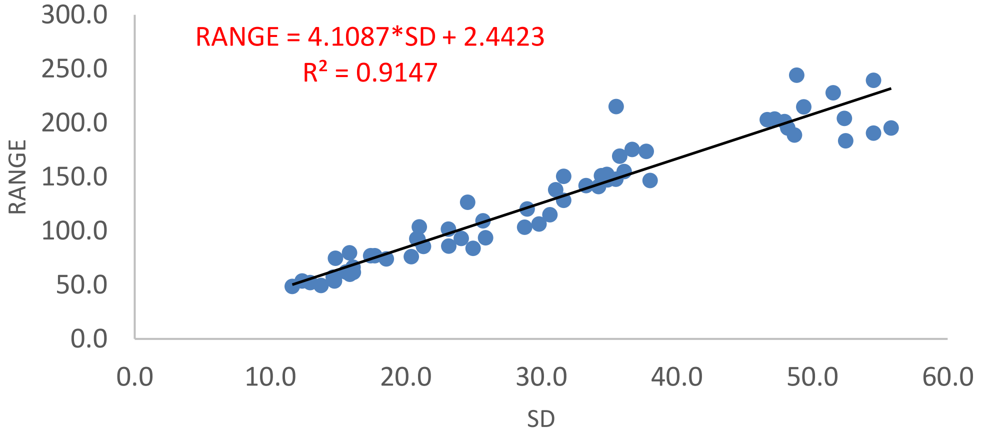


Fig. 2: The Relationship Between the Range and the SD When the Sample size is 30 (n =60)



The relationship between the Range and the SD, based on the 60 pairs of data, derived from the 60 Normal samples of size 75, is shown in Fig. 4. The value of R² = 0.9357, indicates that the Range can be estimated easily with the regression equation of the Range and the SD, shown in the Fig. 4. The slope obtained is 4.78 which suggests that it is possible to account for almost all the observations by Mean ± 2.39*SD when the sample size is 75.

Fig. 3: The Relationship Between the Range and the SD When the Sample size is 50 (n =60)

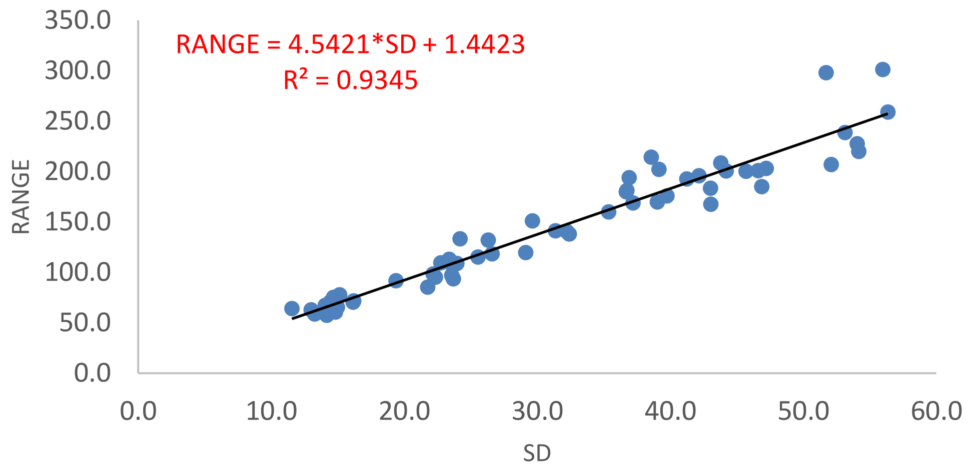
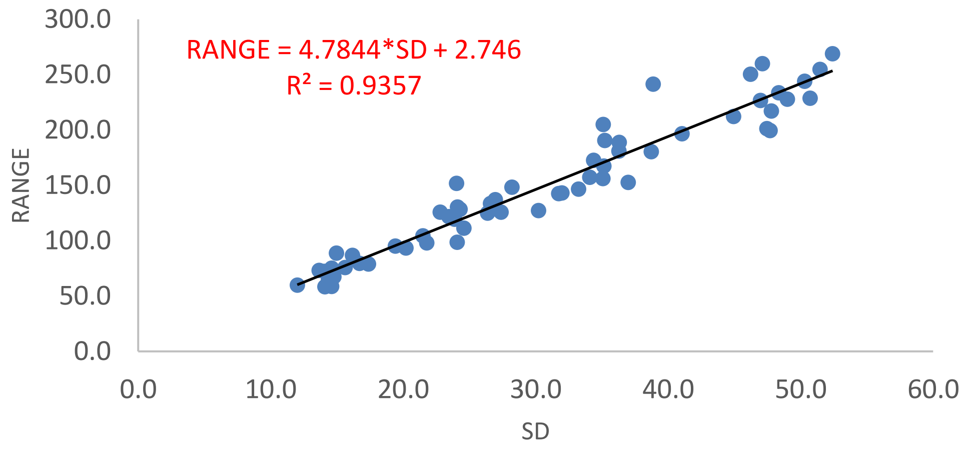


Fig. 4: The Relationship Between the Range and the SD When the Sample size is 75 (n =60)



The relationship between the Range and the SD, based on the 60 pairs of data, derived from the 60 samples of size 100, is shown in Fig. 5. The value of R² = 0.9542, indicates that the Range can be estimated easily with the help of the regression equation of the Range and the SD, shown in the Fig. 5. The slope obtained is 4.98 which suggests that it is possible to account for almost all the observations by Mean ± 2.49*SD when the sample size is 100.

Fig. 5: The Relationship Between the Range and the SD When the Sample size is 100 (n =60)

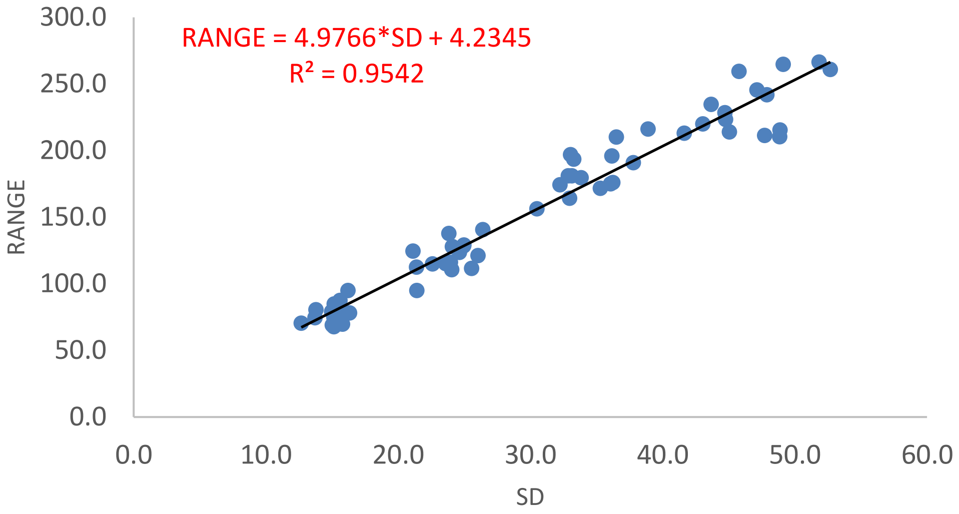
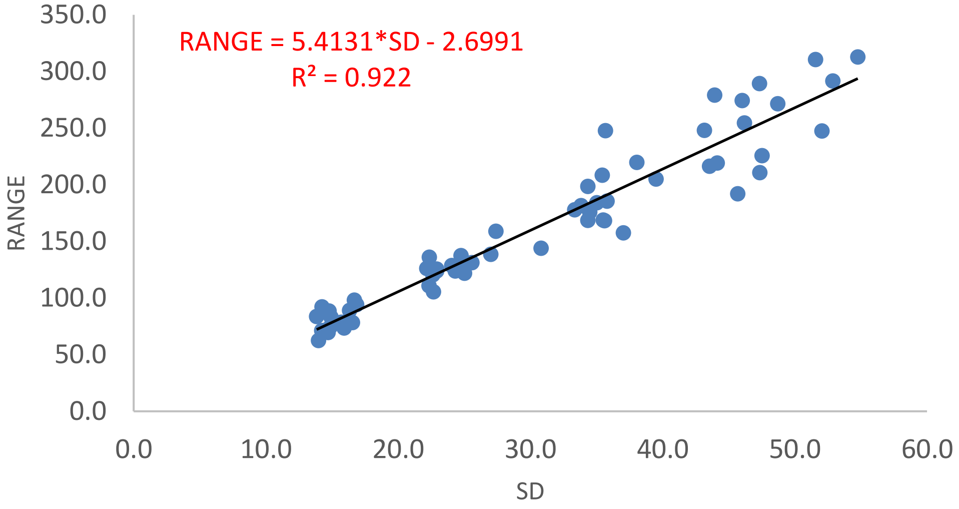


Fig. 6: The Relationship Between the Range and the SD When the Sample size is 125 (n =60)



The relationship between the Range and the SD, based on the 60 pairs of data, derived from the 60 samples of size 125, is shown in Fig. 6. The value of $R^2 = 0.922$, indicates that the Range can be estimated easily with the help of the regression equation of the Range and the SD, shown in the Fig. 6. The slope obtained is 5.41 which suggests that it is possible to account for almost all observations by Mean ± 2.71*SD when the sample size is 125.

The relationship between the Range and the SD, based on the 60 pairs of data, derived from the 60 samples of size 150, is shown in Fig. 7. The value of $R^2 = 0.9505$, indicates that the Range can be estimated easily with the help of the regression equation of the Range and the SD, provided in the Fig. 7. The slope obtained is 5.10 which suggests that it is possible to account for almost all observations by Mean ± 2.55*SD when the sample size is 150.

Fig. 7: The Relationship Between the Range and the SD When the Sample size is 150 (n =60)

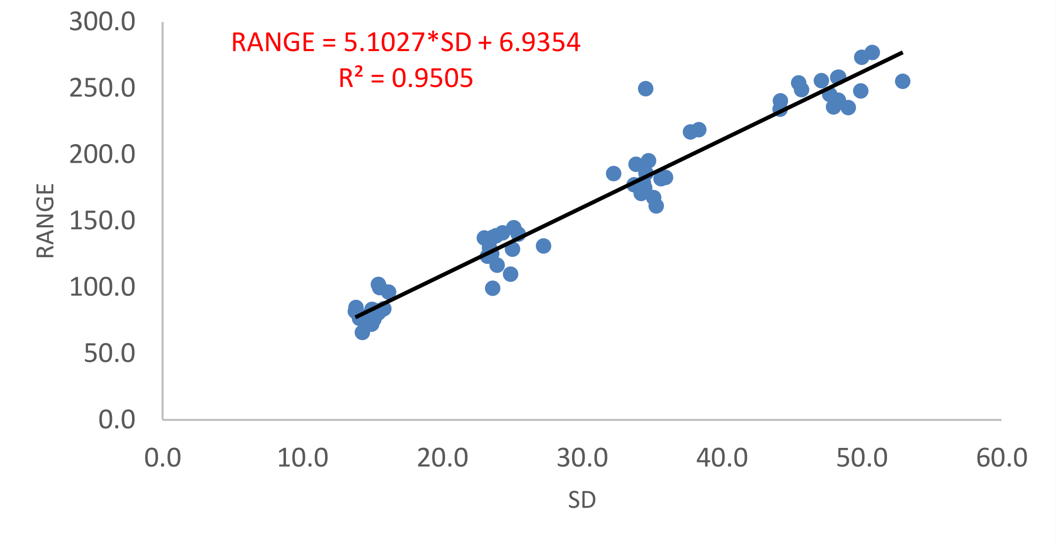
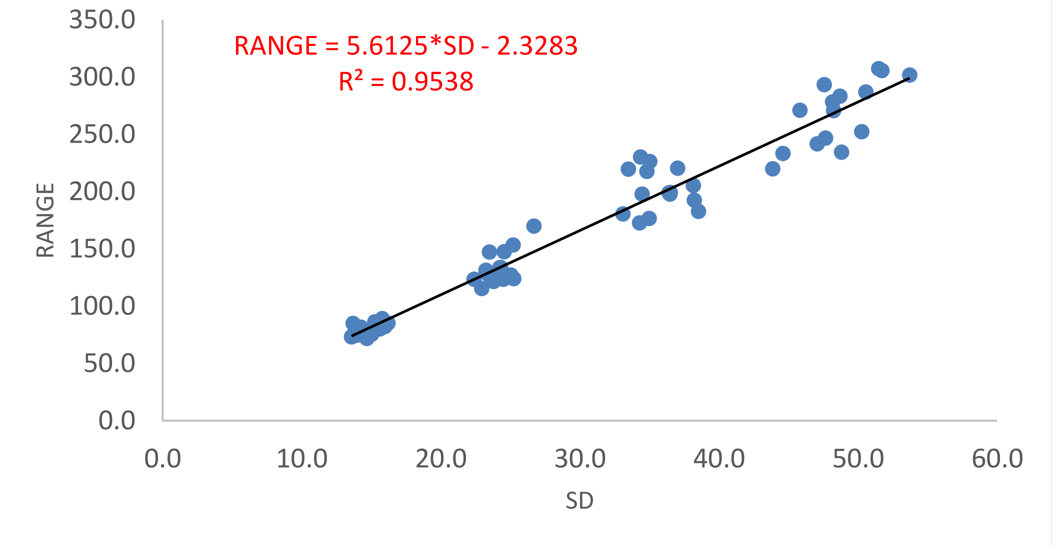


Fig. 8: The Relationship Between the Range and the SD When the Sample size is 200 (n =60)

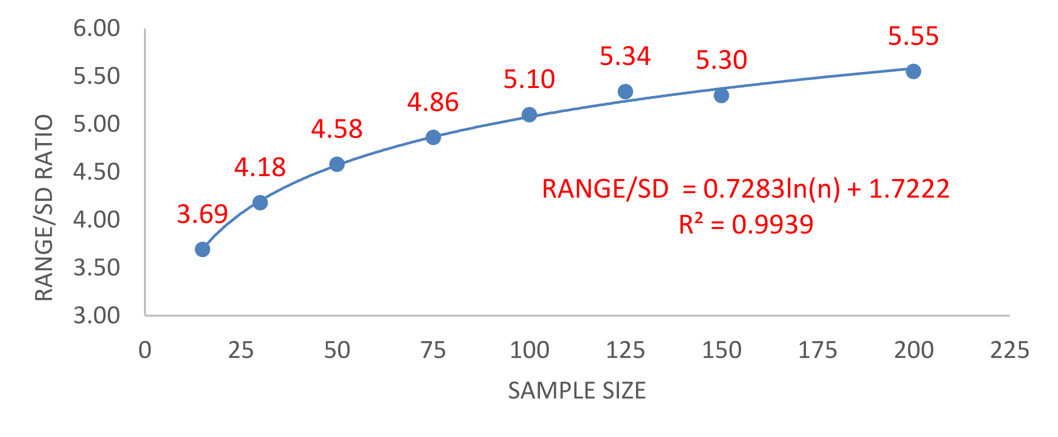


The relationship between the Range and the SD, based on the 60 pairs of data, derived from the 60 samples of size 200, is shown in Fig. 8.

The value of $R^2 = 0.9538$, indicates that the Range can be estimated easily with the help of the regression equation of the Range and the SD, shown in the Fig. 8. The slope obtained is 5.61 which suggests that it is possible to account for almost all the observations by Mean ± 2.81*SD when the sample size is 200.

SAMPLE SIZE AND THE RANGE/SD RATIO

To see the possible effect of sample size on the Range/SD ratio, the slopes are reobtained by setting the intercept as 0 and are plotted against the sample size. The XY plot is shown in the Fig. 9. The R^2 value of 0.9939 suggests that the model fitted is good and given the SD, the Range can be estimated easily by the Regression equation provided in the figure.



It is to be noted that the slope for the sample size of 15 is 3.69 which increases to 5.10 for the sample size of 100 and 5.55 for the sample size of 200, thus, suggesting that there is a positive correlation between the sample size and the Range/SD ratio. From the regression equation, we get

Range/SD = 0.73*ln(n) + 1.72.

⇨ Range = SD*(0.73*ln(n) + 1.72). ……………………… (1)

⇨ $\frac{Range}{2} = \frac{\text{SD}*(0.73*\ln(n) + 1.72)}{2}$

⇨ $\frac{Range}{2}$ = SD*(0.37*ln (n) + 0.86)

Assuming that the Range is equally distributed about the mean, the Low fence and High fence values for determination of the Outliers can be written as follows:

Low fence = LF = Mean - $\frac{Range}{2}$ = Mean - SD*(0.37*ln (n) + 0.86)

High fence = HF = Mean + $\frac{Range}{2}$ = Mean + SD*(0.37*ln (n) + 0.86)

EXTERNAL VALIDITY OF THE REGRESSION EQUATION IN ESTIMATING THE RANGE

The model developed based on the data analysis showed that the Range can be estimated with the greater confidence, as evident by the R^2 = 0.9939. It is natural that the model fitted will be good for the data on which it is developed, however, it is always desirable to test the validity of such a model developed on some other external data.

The four sets of the data generated with specified Mean and the SD (Table 3) are subjected to testing of the validity of the model developed to estimate the Range. The actual Range is compared with that of the Range estimated with the help of the developed model. The absolute difference between the actual Range and the estimated Range value, expressed as a percentage, is termed as the % Error and classified in five class intervals and shown in the Fig. 10. It can be seen from the Fig. 10 that based on the pooled data, more than 85% of the Estimated Ranges are within 15% of the actual values suggesting that the model is good enough for the further application.

THE COMPARISONS OF OUTLIERS BY THE MEAN-SD AND IQR METHOD

For the identification of the Outliers, by IQR and Mean-SD method, Eight Normal samples are generated with specified mean and SD and are shown in Table 4.

Fig. 10: Percentage Error in Range Estimation by the Linear equation (n = 40) For the Sample size of 10, 20, 40,80

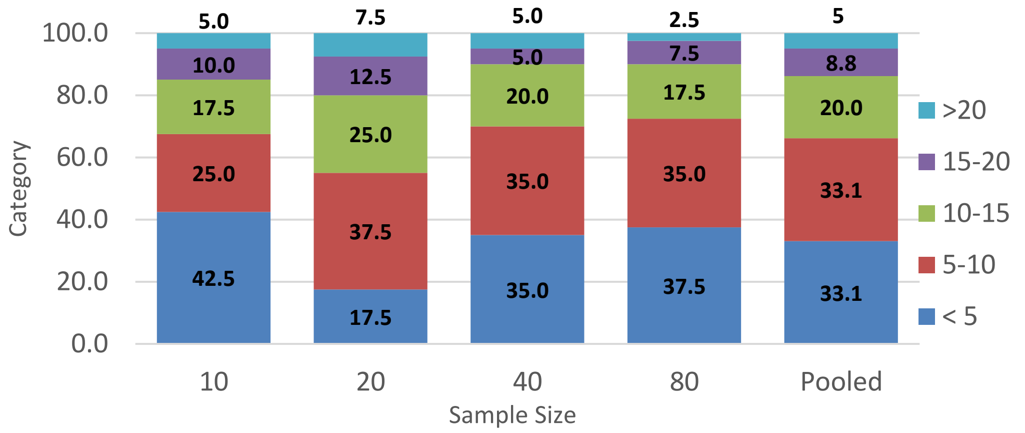


Table 4: Normal Samples for the Identification of the Outliers

Sl. No.	S1	S2	S3	S4	S5	S6	S7	S8
1	25.12	23.03	35.80	31.77	34.71	16.14	27.38	31.04
2	45.64	43.61	38.90	46.82	62.64	56.47	32.74	37.68
3	55.68	48.67	48.96	57.06	71.24	76.23	48.60	44.71
4	55.87	49.32	49.25	62.13	75.03	83.39	63.66	67.48
5	59.80	50.04	51.60	68.62	79.53	103.52	63.92	68.26
6	60.99	55.39	51.99	79.93	102.85	106.72	88.06	75.95
7	68.59	60.08	54.68	86.56	111.60	108.73	101.46	78.28
8	69.66	60.88	56.15	93.44	118.01	110.98	102.59	89.22
9	73.45	62.44	58.33	103.22	119.58	115.14	105.44	91.31
10	74.21	63.29	69.46	104.01	120.75	126.96	108.23	106.92
11	75.04	64.75	70.21	104.62	123.85	132.55	126.63	109.44
12	76.22	64.82	74.34	111.29	125.14	137.86	139.75	113.60
13	76.81	69.26	78.54	112.84	130.21	141.57	148.78	120.58
14	78.37	69.26	87.10	113.01	140.16	166.69	150.97	122.12
15	80.74	72.50	90.03	119.63	145.08	186.61	169.71	142.71
16	84.35	73.80	90.49	124.78	153.92	186.75	170.76	148.34
17	85.74	81.10	96.64	126.81	157.09	213.25	175.85	150.90
18	92.34	82.55	97.47	127.81	170.04	213.32	211.85	154.90
19	92.80	86.01	111.71	128.96	184.81	217.22	217.29	188.85
20	108.32	112.26	120.97	177.99	211.27	217.45	249.25	220.22

For the samples shown in Table 4, the Lower Fence and Higher Fence values are calculated by both the methods of IQR and Mean-SD method and are shown in Table 5. Besides, the Size of the Fences, IQR Relative Size is also calculated and provided. It is surprising to note that 4 out of 8 LF values, in the case of IQR, are negative and are found to be -5.49, -15.17, -31.10, -49.89. In comparison, all the corresponding LF values are positive for the SD-RANGE method. In general, the Relative Size of the IQR method is observed to be higher, ranging from 1.06 to 1.64 times.

Table 5: Lower Fence, Higher Fence, and size of the fence by IQR and SD-Range Method

METHOD	STATISTIC	NORMAL SAMPLES FOR IDENTIFICATION OUTLIERS							
		S1	S2	S3	S4	S5	S6	S7	S8
SD-RANGE METHOD	MEAN	71.99	64.65	71.63	99.07	121.88	135.88	125.15	108.13
	SD	17.87	18.04	23.65	33.38	42.21	55.74	60.87	48.49
	C	35.18	35.51	46.56	65.70	83.10	109.71	119.81	95.45
	LF	36.81	29.14	25.07	33.37	38.78	26.17	5.33	12.67
	UF	107.16	100.16	118.19	164.76	204.97	245.59	244.96	203.58
	SIZE*	70.35	71.02	93.12	131.39	166.19	219.43	239.62	190.91
IQR METHOD	Q1	60.70	54.05	51.89	77.11	97.02	105.92	82.02	74.03
	Q3	81.64	72.82	90.15	120.92	147.29	186.65	169.97	144.12
	IQR	20.95	18.77	38.26	43.81	50.27	80.73	87.95	70.09
	LF	29.27	25.89	-5.49	11.39	21.62	-15.17	-49.89	-31.10
	UF	113.07	100.99	147.53	186.64	222.69	307.74	301.89	249.25
	SIZE*	83.79	75.10	153.03	175.25	201.07	322.91	351.78	280.35
IQR- RELATIVE SIZE**		1.19	1.06	1.64	1.33	1.21	1.47	1.47	1.47

* UF - LF; ** - Size ratio of the SD-RANGE Method and IQR Method e.g., 83.79/70.35 = 1.19

IDENTIFICATION OF OUTLIERS IN SELECTED SAMPLES

The Outliers are identified by both the methods and shown in Table 6. According to Mean-SD Method, 12 Outliers are identified out of which 6 are also identified as Outliers by IQR Method. It is to be noted that those Outliers which are identified only by the Mean-SD method, but not by IQR method, have Z-values more than 2 showing that they can be justifiably considered as the Outliers.

Table 6: Identification of Outliers by MEAN-SD and IQR Method In Generated Eight Samples

SAMPLE	MEAN-SD METHOD				IQR METHOD			
	O1	Z-VALUE	O2	Z-VALUE	O1	Z-VALUE	O2	Z-VALUE
S1	25.12	-2.62	108.32	2.03	25.12	-2.62		
S2	23.03	-2.31	112.26	2.64	23.03	-2.31	112.26	2.64
S3	120.97	2.09	-	-	120.97	2.09	-	-
S4	31.77	-2.02	177.99	2.36	-	-	-	-
S5	34.71	-2.06	211.27	2.12	211.27	2.12		
S6	16.14	-2.15	-	-	-	-	-	-
S7	249.25	2.04	-	-	-	-	-	-
S8	220.22	2.31	-	-	-	-	-	-

DISCUSSION

The study has considered sixty samples each for the sample size of 15, 30, 50, 75, 100, 125, 150 and 200. For each sample size, the SD and the Range are obtained and the relationship was explored. The Range to SD ratio varies from 3.69 for the sample size of 15 to 5.10 for the sample size of 100 and 5.55 for the sample size of 200. In each case, R^2 is observed to be more than 0.90 which can be considered as an index of strong relationship between the SD and the Range. This strong relationship led to the development of a model by which the linear relationship between the SD and the Range, involving the sample size n, is shown below:

Range = SD*(0.73*ln(n) + 1.72). or Range/SD = 0.73*ln(n) + 1.72

This model clearly shows the effect of the sample size on the relationship between the SD and the Range. It is further logical to think that if a model is developed on certain data, it certainly holds good for that data. In model development, one must be careful about the likely presence of the confounding factors. Sometimes, due to presence of some confounding factors the relationship between any two variables may appear to be good but when it is applied to other sets of data, it may not yield the anticipated good results. In view of this, going for the external validity of the model developed, is always considered as a golden rule. Once the model is proved to be good on the external data, the data other than on which the model is developed, the model may be deemed to be a good model for further applications.

It is to be noted that not only the developed model shows the relationship between the Range and the SD, it also allows you to estimate the Range, given the SD and the sample size. When the model developed is applied to four other sets of data, it has shown that in around 65% of the cases, the Range estimated using the model is within 10% of the error and around 85% of the cases, the error is within 15% from the actual Range. This establishes the external validity of the model and allow us to claim that it is suitable for application to any external data.

The estimation of Range from the SD of the data, also give us an opportunity to develop an equation which would be helpful in identification of the Outliers in the data. In literature, IQR use in identification of the Outliers is available. The use of the equation developed, in terms of mean and SD, to identify the Outliers from eight sets of data, showed that the equation can pick up 12 Outliers while in comparison, IQR approach identifies only 6 Outliers that is 50%. This may mean, first that the equation developed in this study is a superior method in identification of the Outliers and Second, that the equation developed is picking up unnecessarily a greater number of Outliers than the picked up by the IQR approach. When those extra picked up Outliers are converted to Z values, it is observed that all those points have value more than 2, justifying the labelling of those points as Outliers. This confirms that the newly developed approach in this study is superior to IQR approach, in identification of the Outliers.

SUMMARY OF THE OBSERVATIONS

- There is a definite relationship between the SD and the Range of the data.
- The relationship is a function of the sample size therefore may keep changing with change in the sample size.
- The equation developed is Range = SD*(0.73*ln(n) + 1.72).
- The establishment of the External Validity of the model proves that it is suitable for application to any external data
- A new method involving the Mean and SD is developed for identification of the Outliers.
- Assuming that the Range is equally distributed about the mean, the Low fence and High fence values for determination of the Outliers can be written as follows:

  Low fence = LF = Mean - $\frac{Range}{2}$ = Mean - SD*(0.37*ln (n) + 0.86)

  High fence = HF = Mean + $\frac{Range}{2}$ = Mean + SD*(0.37*ln (n) + 0.86)

- The method developed has shown to be better in picking up of the Outliers, 50% more, as compared to the that picked up by the IQR method.
- The IQR assign, on an average, 35% higher fence size, as compared Mean SD Method.
- IQR method also include the negative values in their fence values which is surprising as all the data are of positive values.
- There is a need to check the validity of 1.5 as the multiplier of IQR to achieve the lower and higher fence values.

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]. Frost J 2023: Range Rule of Thumb: Overview and Formula, Statistics by Jim https://statisticsbyjim.com/basics/range-rule-of-thumb/

[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]. Schenkelberg F 2023: The Range Rule, Accendo Reliability https://accendoreliability.com/the-range-rule/

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

[8]. StatPlus 7.6.5.0 2021, AnalystSoft Inc. - Statistical analysis program. Version v7. See https://www.analystsoft.com/en/

[9]. Taylor C 2023: Range Rule for Standard Deviation, Thought Co https://www.thoughtco.com/range-rule-for-standard-deviation-3126231

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 11 research papers during 2021-23.