Concept of Outlier Study: The Management of Outlier Handling with Significance in Inclusive Education Setting

Abikesh Prasada Kumar Mahapatra[1*], Anita Nanda[2], Bibhuti Bhusan Mohapatra[2], Archana Kumari Padhy[2] and Indira Padhy[3]

[1]Opjs University, School of Pharmacy, Churu, Rajasthan-331303, India.
[2]DPMIASSE&T, Special BEd. College, Kodola, Ganjam, Odisha-761032, India.
[3]Khallikote Autonomous College, Berhampur, Ganjam, Odisha-760001, India.

Authors' contributions

We declare that this work was done by the authors named in this article. Authors APKM and AN conceived and designed the study. Authors AKP and IP wrote the protocol, managed the literature searches and collected the information related to article. Author AN drafted the manuscript. Author APKM supervised the work and assisted in the final drafting. Author BBM performed the statistical analysis. Authors AN and BBM contributed to final revision of the manuscript. All authors have read and approved the final manuscript.

Case Study

Abstract

Collection of data and to check its suitability is the first step in any statistical data analysis. In such analyses, the presence of outliers appears as an unavoidable important problem. Outliers are unexpected random values in dataset, and they can alter the statistical conclusion and also affect their assumptions. Thus, in order to manage the data properly, outliers must be defined and treated. So all statisticians have to confront the analysis and forced to take a decision. There is only being one of the two extreme choices left for the researcher or statistician during the analysis of outliers. First, either to reject the outlier with the risk of loss of genuine information and the second one is to include them with the risk of error in drawing conclusion. The study therefore summarize the various potential causes of extreme scores in a data set (e.g., data recording or entry errors, sampling errors, and legitimate sampling), how to detect them, and whether they should be removed or not. Another objective of this study was to explore how significantly a small proportion of outliers can affect even simple analyses. The study was explored with

*Corresponding author: E-mail: abikeshmahapatra@gmail.com;

citing suitable examples including outlier value and also excluding the outlier data. The examples show a strong beneficial effect of repetition of the study based on extreme of scores. One way ANOVA test was performed and the significance of extreme outlier was described.

1 Introduction

Various scientific disciplines regularly come across 'outliers' in their data [1,2]. Outliers are commonly defined as observations which are different from the majority of other cases in a sample [3,4]. In an outlier study it is important to identify whether the observation is a genuine member of the main population or it is a contaminant from a different source. The contaminant may lead to incorrect inferences about original population. Sometime the presence of contaminants is not significant, thus the inference process is not effect significantly. When there is a weird observation in the studied data, then to draw a possible inference is very difficult. Such divergence in a population data could be due to, measurement or coding error, sampling from the wrong population, exceptional circumstances, or a poor fit of the statistical model. Before starting the process the principle mass of data, the data must be examined for the presence of outliers. The important aspect of the outlier study is that whether it is accepted or rejected. The answer is mainly depends on the studied population data. Outliers matter and there are several famous cases where even a single case heavily impacts the results [5].

Despite the importance of outliers, researchers do not have clear guidelines about how to deal with them properly. Furthermore, although in many cases outliers are seen as ''data problems'' that must be ''fixed,'' outliers can also be of substantive interest and studied as unique phenomena that may lead to novel theoretical insights [6]. Thus, there is a need for a better understanding and clear guidelines regarding the following three issues: (a) how to define them (i.e., ''what exactly is an outlier?''), (b) how to identify them (i.e., ''How do I know whether a particular case is an outlier?''), and (c) how to handle them (i.e., ''What do I do with a case that has been identified as an outlier?''). At present, researchers are faced with multiple and often conflicting definitions of outliers, techniques to identify outliers, and suggestions on what to do with outliers once they are found [7,8].

It is thus clear that outliers are often reported and that a commonly applied procedure is to remove them. The present work investigated how often statistical conclusions based on 'P = 0.05 significance' diverged when a test with outlier exclusion yielded a statistically significant result whereas the test with outlier inclusion did not, or vice versa.

In the present case study the difference in P value was analyzed when these statistical conclusions are differs due to outlier. Finally, the present study also investigates how the test results effected when reporting any result that was statistically significant (either the test with inclusion or exclusion of outliers). The current research utilized ANOVA design, studied the P value and simulated the result with respect to presence or absence of outlier. The data obtained with significant outlier was again repeated for duplicate study and the result was compared with the previous data.

2 The Concept of Outlier Study

2.1 The myth about outlier study

More often the concept of outlier and noise are considered similar. However, the concept of outlier is essentially different than noise. While Outlier is a data that significantly different compared to the other data, noise is a random error or variance. The outlier is part of the data, but Noise is just a random error (could be

mislabeled or mistake or even missing data). Many parametric statistics, like mean, correlations, and every statistic based on these is sensitive to outliers. Since the assumptions of standard statistical procedures or models, such as linear regression and ANOVA also based on the parametric statistic, outliers can mess up analysis [9].

2.2 Importance of outlier study

An outlier is an observation that appears to deviate markedly from other observations in the sample.

Identification of potential outliers is important for the following reasons [10].

1. An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).
2. In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any case, typically the outlying observation is not simply deleted. However, if the data contains significant outliers, then there is requirement of consider the use of robust statistical techniques.

2.3 Problems and assumptions to be addressed while dealing with outliers

2.3.1 Problems associated with outlier study

There are many issues observed while dealing with data with regards to outliers. As discussed with Iglewicz and Hoaglin (1993), there are mainly three types of issues associated with outliers [11].

1. **Outlier labeling** – To find out the reason of observed outlier and to investigate whether the potential outlier is an erroneous data or cause of an inappropriate model selection or so on.
2. **Outlier accommodation** – Selection of robust statistical techniques that will not affected by outliers. If the observed potential outliers are incorrect observation, then to take decision whether there is need of modifying the statistical method or to accept the observation.
3. **Outlier identification** – To identify the outlier by formal test.

2.3.2 Assumptions related to outlier study

Identifying the observation of an outlier in data set solely depends on the distribution of the data. It is highly essential that the data set to follow univariate model i.e. follow an approximately normal distribution. If the normality of the data is not tested before outlier identification then it is very much difficult to distinguish whether the outlier is due to non normality of data or due to some other probable reasons [10]. For this reason, it is advisable to study the distribution pattern of the data and to generate a normal probability plot before applying outlier test. Even if the normality test is performed, in many instances, the presence of one or more outliers may cause the tests to reject normality when it is in fact a reasonable assumption for applying the outlier test. The box plot and the histogram can also be useful graphical tools in checking the normality assumption and in identifying potential outliers.

2.4 Preliminary requirement of outlier test

2.4.1 Data considerations for outlier test

To ensure the obtained results are valid, following guidelines should be considered while collecting data, performing the analysis, and interpreting the results [12,13].

1. The data must be numeric

The data must have numeric data, such as the weights of packages.

2. The sample data should be selected randomly

In statistics, random samples are used to make generalizations, or inferences, about a population. If the data are not collected randomly, the results may not represent the population.

2.4.2 Single vs. Multiple outlier

The statistical methods selected are based on the type of outlier. There are some methods which can detect a single outlier where as some statistical tests are there which can detect multiple outliers present in the population. It is not advisable to use a statistical test appropriate for detecting a single outlier sequentially in order to detect multiple outliers [14,15].

2.5 Causes of outlier

It is very important to decide whether the outlier from the data should be removed or not. If it is not possible to remove then the process of dealing the outlier in the dataset is a critical aspect for evaluation. For these types of cases the proper action depends on what causes the outliers.

In general, there are three causes for outliers— (1) data entry or measurement errors, (2) sampling problems and (3) unusual conditions, and natural variation [14].

(1) Data entry and measurement errors

Errors can occur during measurement and data entry. During data entry, typos can produce weird values. For example marks of 10 students were measured and collected the following dataset.

Example of Data entry and measurement errors causing an outlier

Student and their marks in a class room setting

Students	1	2	3	4	5	6	7	8	9	10
Marks (%)	55	60	311	63	58	62	57	62	59	64

In the above dataset, the value of 311% is clearly an outlier. Not only does it stand out, but it's an impossible mark value. Examining the numbers more closely, it is understood that the one might have been accidental. Hopefully, in this case there is requirement of to check the original record. These types of errors are easy cases to understand. If it is understood that an outlier value is an error, then the value was corrected when possible. That can involve correcting the typo error or possibly reevaluating the mark of particular student. If that's not possible, you must delete the data point must be deleted as we know that it is an incorrect value.

(2) Sampling Problems or unusual conditions

Inferential statistics are always used to draw conclusion or inference for a specific population. For a population the study was usually conducted by drawing a random sample from the define population. Unfortunately, if the random item or person accidentally selected from other population rather than the target population then it shows a wrong inference about the studied population. There are several ways this can occur. For example, unusual events or characteristics can occur that deviate from the defined population. In

other cases, an item was accidentally collected that falls outside the target population, and, thus, it might have unusual characteristics.

Examples of sampling problems causing an outlier

There is manufacturing defect in a production unit due to abnormal function of the equipment. However, the outliers product obtained assumed may be due to normal functioning of equipment and abnormal function of equipment is ignored. Products manufactured under these unusual conditions of equipment do not reflect in target population of products from the normal process. Consequently, it is better to legitimately remove these data points from the dataset.

(3) Natural variation

All data distributions have a spread of values. Extreme values can occur, but they have lower probabilities. If the sample size is large enough, it is obvious to obtain unusual values. In a normal distribution, approximately 1 in 340 observations will be at least three standard deviations away from the mean. However, random chance might include extreme values in smaller datasets.

Example of natural variation causing an outlier

Analyzing the performance of a class based on their mark. The statistician applied a model and analyzes the performance of the class based on the fit of model. However, it is observed that one data point (student) severely affects the model. If the data set removed the class performance is good but if it is kept in the model the class performance is bad. In this case the removal of one data point significantly affected the inference. Thus, it is recommended that even though this unusual observation is influential, it is better to left as such in the model. It is bad practice to remove data points simply to produce a better fitting model or statistically significant results. If the extreme value is a legitimate observation that is a natural part of the population then it is better to include the particular data in the dataset [14,16].

2.6 Basic approaches to distinguish outlier in data

There are a variety of ways to find outliers. All these methods employ different approaches for finding values that are unusual compared to the rest of the dataset. The assessments tolls used to finding the outlier values in data is broadly divided into two categories (1) visual assessments (2) analytical assessments [14,17].

1. Sorting the datasheet to find outliers
2. Graphing data to identify outliers
3. Using Z-scores to detect outliers
4. Using the interquartile range to create outlier fences
5. Finding outliers with hypothesis tests

2.7 Strategy for dealing with outliers

Sometimes it's best to keep outliers in the data. They can capture valuable information that is part of study area. Retaining these points can be hard, particularly when it reduces statistical significance. However, excluding extreme values solely due to their extremeness can distort the results by removing information about the variability inherent in the study area.

When considering whether to remove an outlier, there is need to evaluate if it appropriately reflects the target population, subject area, research question, and research methodology. There is requirement of analyzing the reason for such type of unusual observations [14]. Some example of these types of observations like

1. Anything unusual happen while measuring these observations, such as power failures, abnormal experimental conditions, or
2. Anything else out of the norm
3. Is there anything substantially different about an observation, whether it's a person, item, or transaction?
4. Did measurement or data entry errors occur?

If the outlier in question is:

I. A measurement error or data entry error, correct the error if possible. If it is not possible to fix it, it's better to remove that observation because it is known that the value is incorrect.
II. Not a part of the population studying (i.e., unusual properties or conditions), remove the outlier legitimately.
III. If studying a natural part of the population, better to not remove it.

When decision taken to remove outlier, documentation needs to be done with both excluding and including the outlier and proper reasoning should be cited for the excluded data point. It is always wise decision to perform the analysis with and without these observations and discuss the differences. Comparing results in this manner is particularly useful when removing an outlier is unsure and when there is substantial disagreement within a group over this question.

2.8 Application of statistical analyses or tools for outliers

There are several situation exists where the statistician is having very limited options to deal with the outlier. Suppose a situation when there is no option to remove outliers and if the outlier present; they can violate the basis assumption of the statistical analysis. In this case the requirement is to include the outlier without compromising the result. For these types of case studies several statistical tools or analyses available which can help to get rid out of this issues [14,17]. Out of several options available, the most frequently used and user free options are explored below

➤ Nonparametric hypothesis tests are robust to outliers.
  In this test outliers won't necessarily violate their assumptions or distort their results.
➤ In regression analysis,
  The data can be transferred by using a robust regression analysis available.
➤ Bootstrapping techniques
  This technique doesn't make assumptions about normal distributions of data.

3 Case Study for Management of Outlier by Using Statistical Tools in Special Education Setting

A case study was presented for the process of dealing outlier in data set during investigation in a special education setting. The case study includes sample selection, design of study, tools used and step wise process for dealing outlier during collection of data for a study design. The case study explains the sequential process with their inference.

3.1 Selection of sample

The population of the study will consist of 10 intellectual disable special school children with mild mental retardation. All intellectual disable students belong to mild category. The study will be conducted in a inclusive education setting with the help of "BALVIKAS" located near Kodola town of Ganjam district Odisha, INDIA. The study also considers only male student as an integral part of the study. These students were selected based on their Intelligence Quotient (IQ) and between age group 8-15 years.

3.2 Design of study

Research design is defined as a framework of methods and techniques chosen by a researcher to combine various components of research in a reasonably logical manner so that the research problem is efficiently handled. It provides insights about "how" to conduct research using a particular methodology. Every researcher has a list of research questions which need to be assessed – this can be done with research design. The sketch of how research should be conducted can be prepared using research design. Hence, the research study was carried out on the basis of research design.

3.3 Tools used

The tools used are mainly the available tools used by the professionals in the field of statistics and intellectual disability. The performance of the students will be reviewed in terms of score in the light of the specific purpose of the present study after 1 hr, 2 hr and 3 hr and appropriate tools will be identified, developed and administered. The main tool used for the evaluation of this study was statistical software Minitab 17.3.0.

3.4 Scoring and interpretation

Ten students are trained by professional trainers for one week to train the students on a topic related to their professional field. After 7 days the students were tested for their performances and the performance was evaluated in terms of scores or marks obtained in percentage after 1 hr, 2 hr and 3 hr. After record of observation scoring was done in the respective time interval .The numerical score for all students were calculated in percentage and reported.

4 Presentation of the Case Study in Special Education Setting for Management of Outlier Data

As per the study design the scores of the ten students were collected after 1 hr, 2 hr and 3 hr time interval. The details of scores of the students with their mean, minimum and maximum marks obtained are presented in Table 1.

Table 1. Scores/marks (%) obtained by students after 1 hr, 2 hr and 3 hr time interval

Students	Marks/scores obtained (%) by students after 1 hr, 2 hr and 3 hr time interval		
	After 1 Hr	After 2 Hr	After 3 Hr
1	38	38	39
2	47	42	46
3	53	40	53
4	35	33	52
5	51	53	67
6	52	47	66
7	66	54	50
8	50	38	52
9	72	12	58
10	52	48	41
Mean	**51.6**	**40.5**	**52.4**
Min	**35**	**12**	**39**
Max	**72**	**54**	**67**

From the above table it was observed that there is variation exists between the marks obtained by students after 1 hr, 2 hr and 3 hr. The mean of 2 hr study was also having significant difference and on lower side as compared to other time interval studies. It was observed that the 9th number student showed an out of trend data after 2 hr study. The out of trend data observed for student may be due to various reasons. Thus, it is highly essential to find the actual reason of variation to make a valuable decision.

Further, to get more information of the unusual observation the route cause analysis (RCA) was carried out.

4.1 Root cause analysis (RCA)

Root cause analysis (RCA) is a systematic process for identifying "root causes" of problems or events and an approach for responding to them. The root cause of variation was investigated by using statistical analysis by using statistical software Minitab 17.3.0. Individual scores obtained by all ten students afte1 hr, 2 hr and 3 hr were considered for evaluation. The evaluation was carried out by following steps

    1.   Outlier analysis

    A.   Histogram plot to check spread and shape of data
    B.   Normality test to check normal distribution of data
    C.   Box plot to check the centre and spread of data

    2.   Test of outlier by using Grubb's test
    3.   One way ANOVA with considering all time interval i.e. 1 hr, 2 hr and 3 hr
    4.   One way ANOVA with excluding 2 hr time study (considering time interval 1 hr and 3 hr)

4.2 Sequential approach for RCA study

4.2.1 Outlier analysis

  A. Histogram plot

Histogram plot helps to examine the distribution of sample data, including the peaks, spread, and symmetry. It also helps to understand the variation in the data. It assesses how the sample size may affect the appearance of the histogram. The scores obtained by the students are subjected to histogram plot to check their distribution pattern.
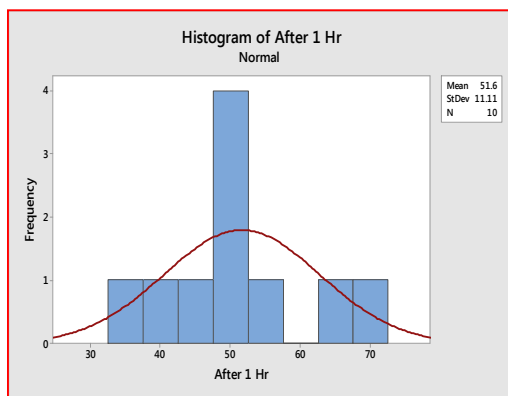

(A)
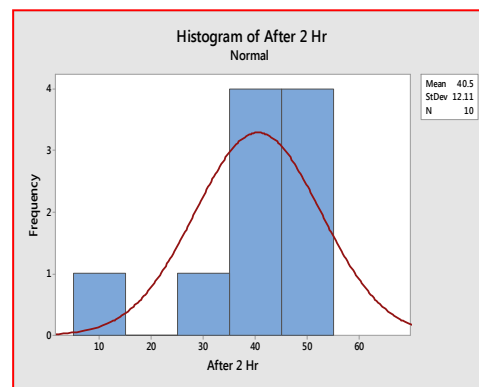
(B)
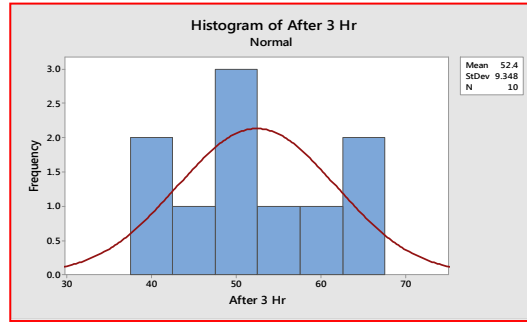
(C)

Fig. 1. Histogram of % score obtained by students (A) after 1 hr and (B) after 2 hr and (C) after 3 hr

From the histogram plot it can be inferred that the scores of 1 hr and 3 hr are symmetrically distributed (Fig. 1A and 1B). However, the scores of 2 hour study are left skewed which is not normal or normal distributed (Fig. 1C). This indicated that there is some abnormality exists in the obtained data of 2 hr study. Thus, for further evaluation the extreme left data point of scores(score of 9th student) obtained by the students after 2 hr was removed and checked.
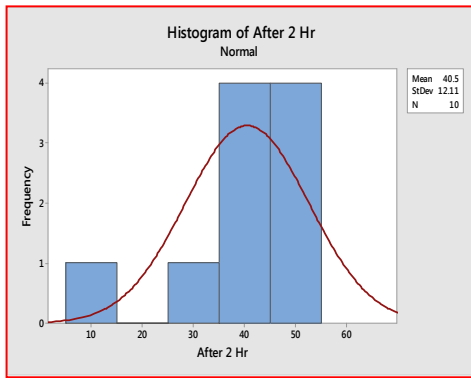

(A)
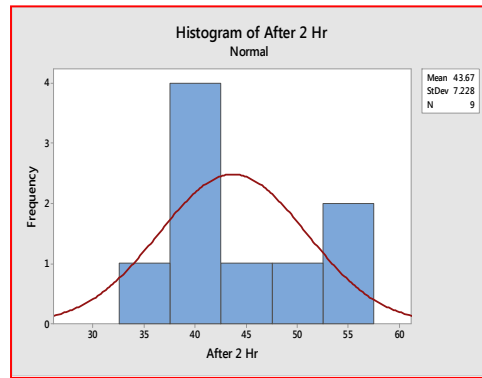
(B)

Fig. 2. Histogram of % marks obtained by students after 2 hr (A) Histogram with (N=10) students and (B) Histogram with (N=9) students: Excluding the extreme data point

From the observed histogram presented in Fig. 2A, it can be seen that the scores of 2 hour time is symmetrical distributed and also follows a bell shaped curve. Based on this it can be concluded that the marks obtained in 2 hour are having unusual observation. Further, to confirm the hypothesis normality test was performed.
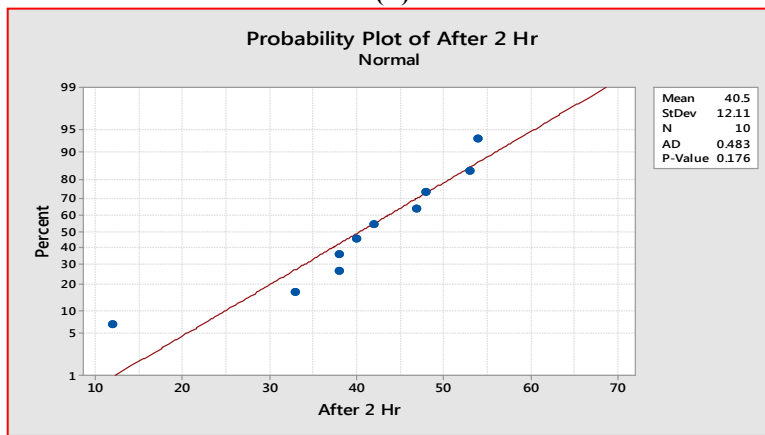
B. Normality test

Normality Test was preferred to check whether the studied data follow a normal distribution or not. For outlier study the main assumption is that the data should follow normal distribution. Thus, it is highly essential to study the distribution of data. The test results of normality test indicate whether to reject or fail to reject the null hypothesis that the data come from a normally distributed population. The normality test and probability plot are the best tools for judging normality. For normality test Anderson-Darling test was used**.** This test compares the ECDF (empirical cumulative distribution function) of the sample data with the distribution expected if the data were normal. If the observed difference is adequately large, the null hypothesis of population normality was rejected. The scores obtained by the students after 1 hr, 2 hr and 3 hr
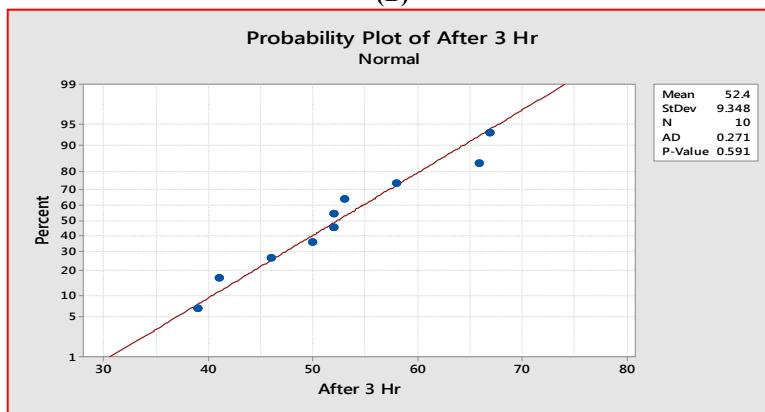
study are subjected to normality test to identify whether the obtained data are from same population or different population.
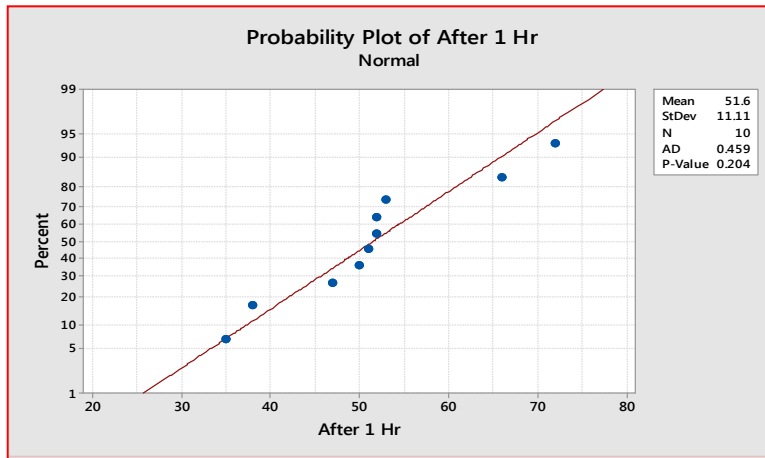
(A)

(B)

(C)

Fig. 3. Normality test of % marks obtained by students (A) after 1 hr and (B) after 2 hr and (C) after 3 hr

The p-value is a probability that measures the evidence against the null hypothesis. Smaller p-values (compared to level of significance 0.05) i.e less than 0.05 provide stronger evidence against the null hypothesis. Larger values for the Anderson-Darling statistic indicate that the data follow the normal distribution. From the Anderson-Darling normality test it was observed that all time intervals having P value more than 0.05 (Fig 3A, 3B and 3C). Thus, it was concluded that the data for all time interval follows normal distribution.

C Box plot study

Box plot was used to assess and compare the shape, central tendency and variability of sample distributions. It helps to identify any outlier present in the data. A box plot shows the median, interquartile range and outliers for each group. In histogram plot it was observed that the 2 hour study follows unsymmetrical distribution and it was also anticipated that it may contains outlier. All prerequisite requirements studies (histogram plot and normality study) were performed for the obtained data. Finally, box plot was performed for 2 hr obtained score to find out the potential outlier.
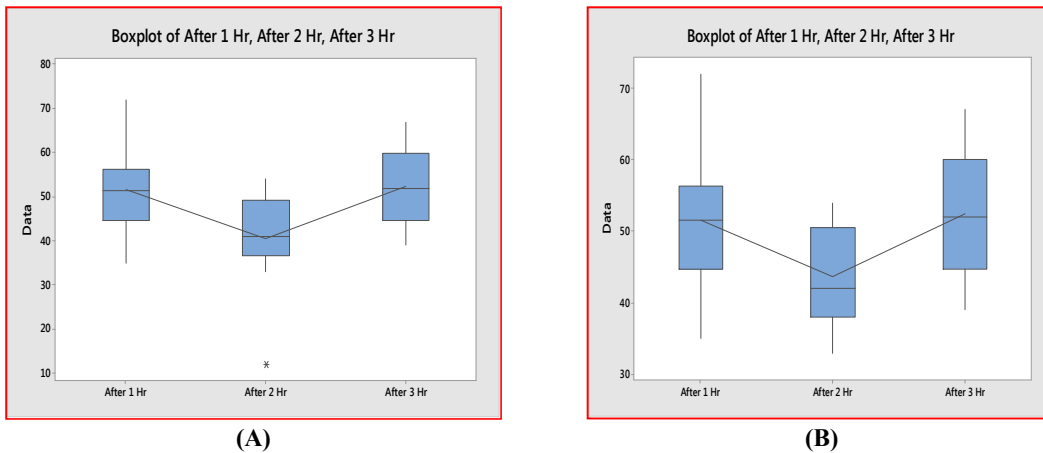

(A)   **(B)**

Fig. 4. Box plot of % marks obtained by students after 1hr, 2 hr and 3 hr (A) Representing the outlier (including 9th observation in 2 hr study) and (B) Excluding the outlier (excluding 9th observation from 2 hr study)

In box plot the data values that are far away from other data values, strongly affect the results. Identification of outliers in box plot is quite easy which are presented by asterisks (*). From the box plot it is observed that a potential outlier (student 9) was present which can be identified by asterisks mark (Fig. 4A). Further, the score of student 9 was removed (Fig. 4B) and the box plot was compared with earlier one. It indicates that the score of student 9 after 2 hr is the potential outlier (asterisk mark absent).

4.2.2 Test of outlier by using Grubb's test

Outlier Test was used to identify a single outlier in a sample. The outlier test considers null hypothesis as all values in the sample are from the same, normally distributed population.

To determine whether an outlier exists, the P-value was compared to the significance level. Usually, a significance level (denoted as α or alpha) of 0.05 works well. A significance level of 0.05 indicates a 5% risk of concluding that an outlier exists when no actual outlier exists.
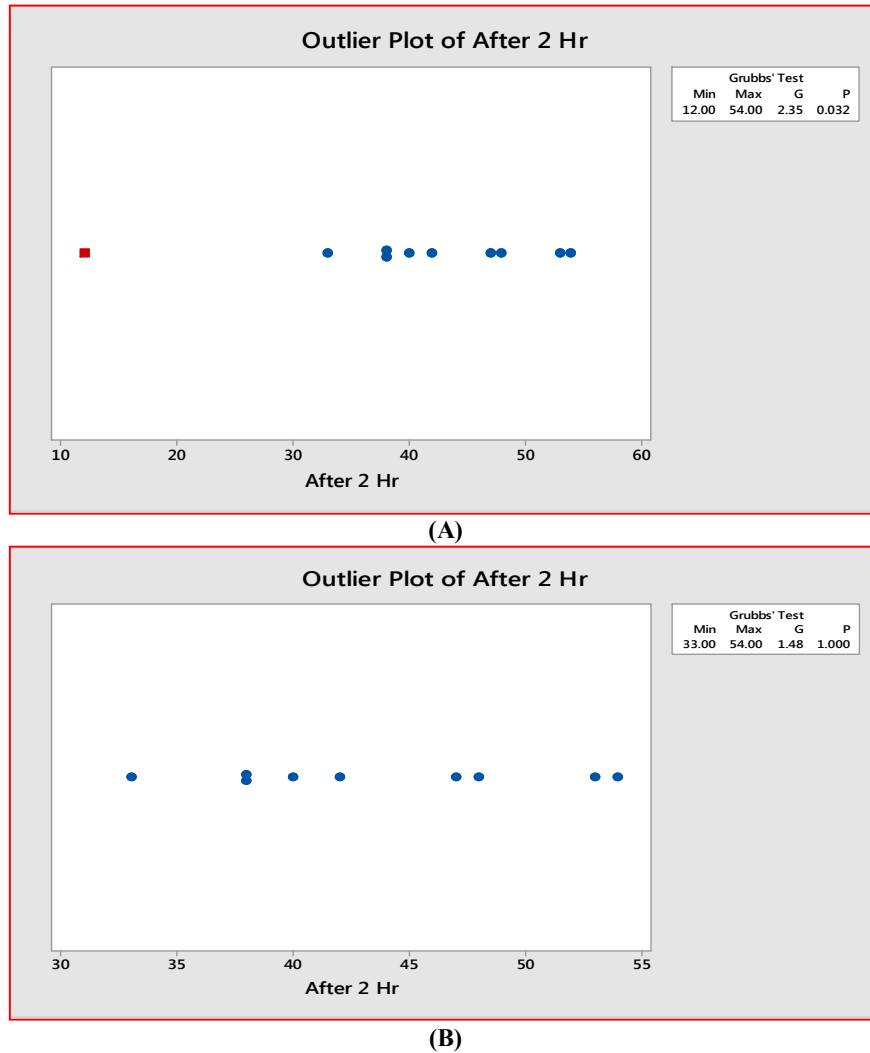
(A)

(B)

Fig. 5. Outlier plot of % marks obtained by students after 2 hr (A) with (N=10) students: Including the outlier and (B) with (N=9) students: Excluding the outlier

In the Grubb's test the null hypothesis states that all the data values come from the same normal distribution. The Grubb's test with the outlier value shows a P value of 0.032 which is smaller than 0.05 significant levels (Fig. 5A). Thus, the null hypothesis rejected. However, when the outlier was removed and test was performed, a P value of 1.000 was observed which is greater than 0.05 significant levels (Fig. 5B). Thus, the null hypothesis accepted and indicated that no outlier in data. Based on this it was confirmed that the observation 9th student in the 2 hr study was an potential outlier.

4.2.3 One way ANOVA for scores considering all time interval i.e. 1 hr, 2 hr and 3 hr

The one-way analysis of variance (ANOVA) is used to determine whether the mean of a dependent variable is the same in two or more unrelated, independent groups of an independent variable. ANOVA is an omnibus test statistic and it helps to identify which specific groups were significantly different from each other; it only helps to determine that at least two groups were different. In the discussed case study one way

ANOVA was applied to identify to know which group is different from the all studied groups (three groups i.e., marks obtained after 1 hr, 2 hr and 3 hr).
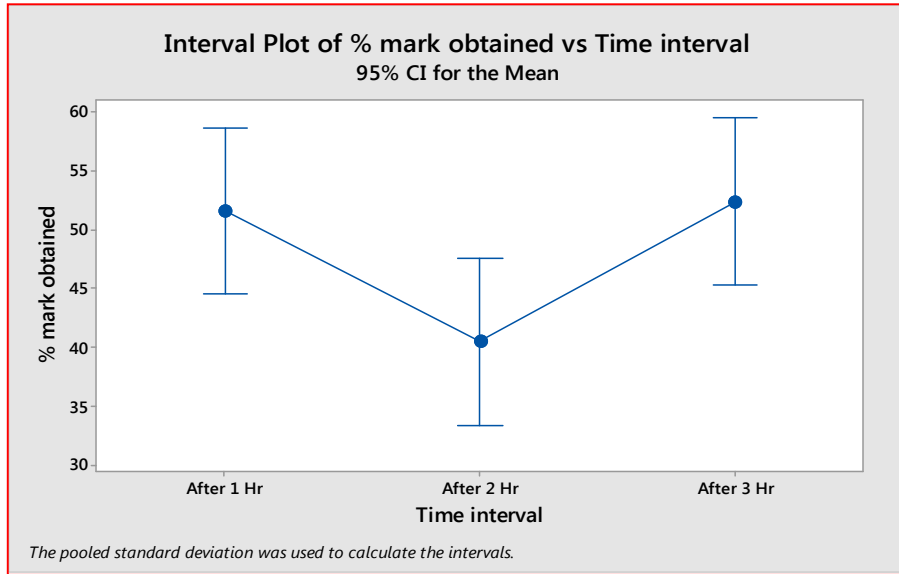

Fig. 6. One way ANOVA plot for marks obtained (%) by students after 1 hr, 2 hr and 3 hr time interval

Table 2. Analysis of variance table

Sr no.	DF	Sum of square	Mean square	F ratio	Prob>F
1.	2	884.9	442.4	3.71	0.038
2.	27	3217.3	119.2		
3.	29	4102.3	-		

Table 3. Means of one way ANOVA table

Duration of study	No. of students	Mean
1 Hour	10	51.60
2 Hour	10	40.50
3 Hour	10	52.40

The result of one way ANOVA reveled that P value was found to be 0.038 which signifies that there is significant difference exist in the scores obtained in different time intervals. The one way ANOVA plot in Fig. 6 also signifies that the data of 2 hr time interval distributed far from the grand mean. For further confirmation, one way ANOVA was conducted by excluding score of 2 hr time interval.

4.2.4 One way ANOVA with excluding 2 hr time study (considering time interval 1 hr and 3 hr)

Further to strengthen the assumption again one way ANOVA was conducted by excluding score of 2 hr time interval study from the population.
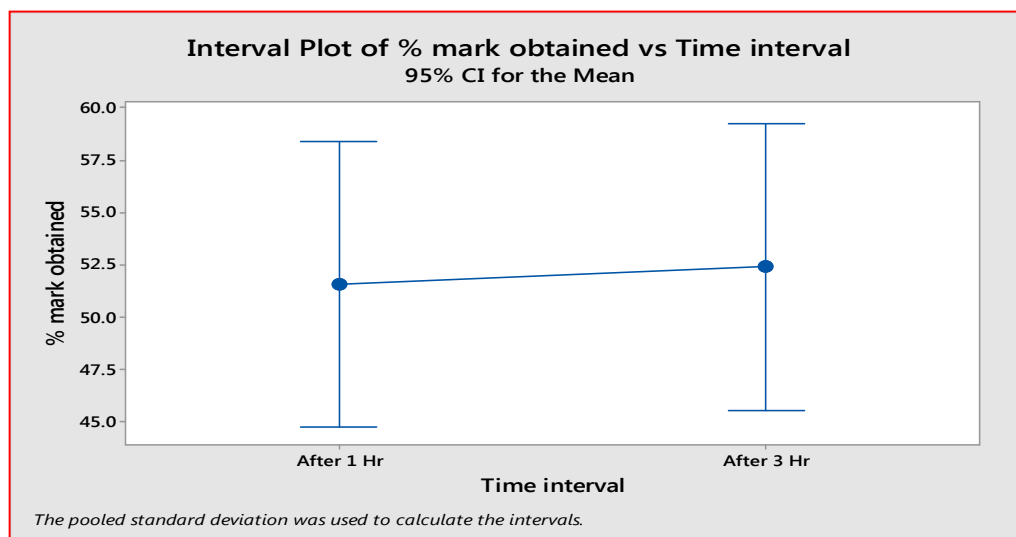
Fig. 7. One way ANOVA plot for marks obtained (%) by students after 1 hr and 3 hr time interval

Table 4. Analysis of variance for marks obtained (%) by students after 1 hr and 3 hr time interval

Sr no.	DF	Sum of square	Mean square	F ratio	Prob>F
1.	1	3.20	3.20	0.03	0.864
2.	18	1896.80	105.378		
3.	19	1900.00	-		

Table 5. Means of one way ANOVA for marks obtained (%) by students after 1 hr and 3 hr time interval

Duration of study	No. of students	Mean
1 Hour	10	51.60
3 Hour	10	52.40

Results of one way ANOVA considering two time interval scores 1 hr and 3 hr revealed that the P value was found to be 0.864 which signifies that there is no significant difference in the scores between 1 hr and 3 hr time interval. The one way ANOVA in Fig. 7 also signifies that both the time intervals are equally distributed around the grand mean and there was no significant difference between these two time points.

From, this study, the results shows that there is a significant difference in 2 hour scores of students (due to presence of outlier) where as no significant difference in 1 hr and 3 hr time study scores of students. Based on the scores obtained and the statistical output from the outlier plot as well as ANOVA plot, it can be concluded that there is requirement of repeating the 2 hour time period study for the students.

4.3 Repetition study

From the outlier plot and ANOVA study it was concluded that the 9th observation of 2 hr time interval study is an potential outlier. It is also observed that if the included in the study statically difference observed between the three studied groups. Moreover, if it is excluded it is observed that the three populations are same. Based, on this observation it was decided that there is requirement of repetition of the study. Thus, it was decided to again evaluate the scores of the students after 2 hour. Scores of 1 hr and 3 hr are kept as such and only 2 hr time point was evaluated again. The new scores obtained are tabulated in Table 6.

Table 6. Individual marks obtained by students (%) after 1 hr, 2 hr and 3 hr time interval after repeting the study

Students	Marks/scores obtained (%) by students after 1 hr, 2 hr and 3 hr time interval after repeting the study		
	After 1 Hr	After 2 Hr	After 3 Hr
1	38	41	39
2	47	46	46
3	53	45	53
4	35	55	52
5	51	50	67
6	52	51	66
7	66	48	50
8	50	44	52
9	72	36	58
10	52	54	41
Mean	**51.6**	**47**	**52.4**
Min	**35**	**36**	**39**
Max	**72**	**55**	**67**

Note: Scores of 1 hr and 3 hr time interval are not repeated. Only 2 hr study was repeated due to observed potential outlier
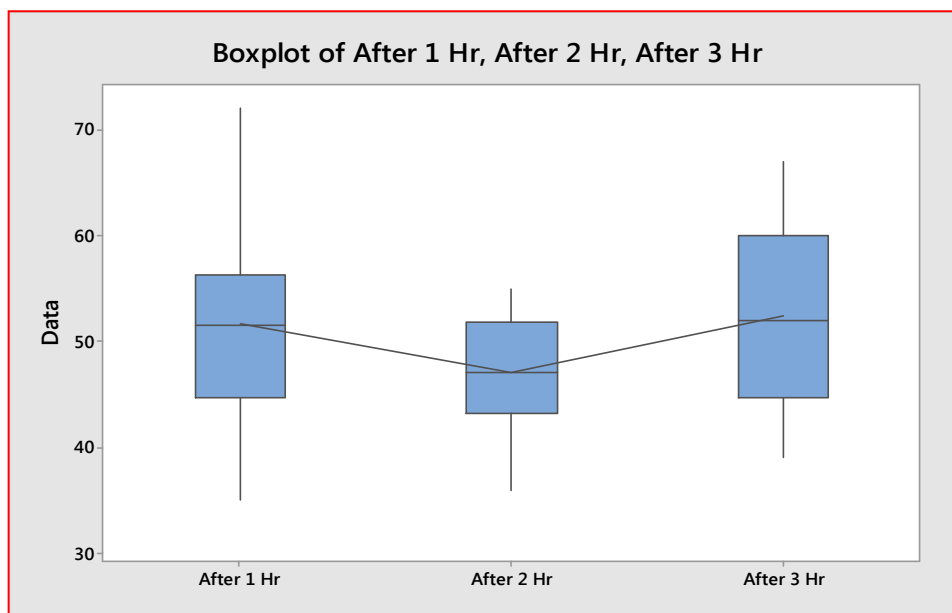

Fig. 8. Box plot of % marks obtained by students after 1hr, 2 hr and 3 hr after repeating the study

The Box plot revealed that there is no existance of outlier (which can be identified by absent of asterisks mark) and the evaluated data found to be close to the grand mean. Further, the score of all students after 2 hr was found satisfactory. There is no indication of outlier data in repeition study as compared to earlier study. The obtained out of trend score observed in the earlier study might be due to the hinderance observed during the inclusive education setting. It happens due to the improper guidance of the special trainer which attributed a poor score and causes a significant difference. In repetiton study outlier test was not performed due to absence of any significant difference between the % of marks obtained by the student after 1 hr, 2 hr

and 3 hr study (Fig 8). ANOVA study was directly performed to find out the whether potetial difference or variation exists between the groups or not.
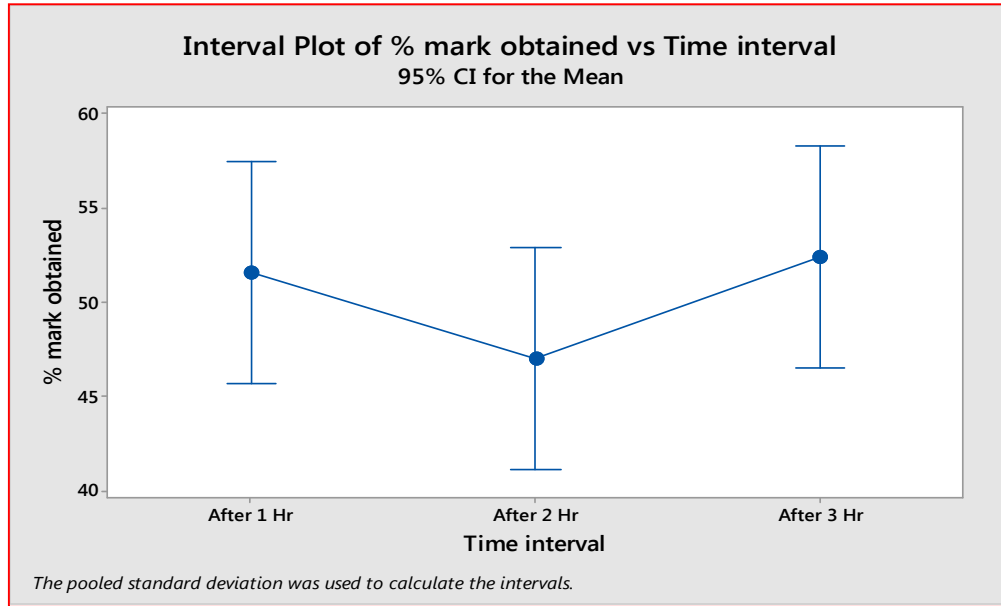
Interval Plot of % mark obtained vs Time interval
95% CI for the Mean

Fig. 9. One way ANOVA plot for marks obtained (%) by students after 1 hr, 2 hr and 3 hr time interval after repeting the study

4.4 ANOVA table summary

An ANOVA test helps to figure out if there is need to reject the null hypothesis or accept the alternate hypothesis. In the present case ANOVA study was performed to identify whether any significant differene exists in the scores of students after 1 hr, 2 hr and 3 hr with repetation of study. The details of ANOVA summary was presented in Tables 7 and 8.

Table 7. Analysis of Variance for marks obtained (%) by students after 1 hr, 2 hr and 3 hr time interval after repeting the study

Sr no.	DF	Sum of square	Mean square	F ratio	Prob>F
1.	2	169.9	84.93	1.04	0.367
2.	27	2206.8	81.73		
3.	29	2376.7	-		

Table 8. Means of one way ANOVA for marks obtained (%) by students after 1 hr, 2 hr and 3 hr time interval after repeting the study

Duration of study	No. of students	Mean
1 Hour	10	51.60
2 Hour	10	47.00
3 Hour	10	52.40

Results of one way ANOVA considering all time interval scores i.e. 1 hr, 2 hr and 3 hr reveals that the P value was found to be 0.367 which signifies that there is no significant difference in the scores between the

time intervals. The one way ANOVA in Fig. 9 also signifies that all the time intervals are equally distributed around the grand mean and there was no significant difference between these two time points. Moreover, significant improvement was observed after repeating evaluation of the data of 2 hr time interval. It was also concluded that the unusual observation observed in the earlier study might be due to the lack of study design, specially the role of special educator in the inclusive education setting.

5 Recommendations

From the repetition study an improvement was observed in the scores of students in 2 hr study. There is no outlier observed in 2 hr study and there is no significant difference observed between the time interval studies. The possible reason of the observed outlier was identified and based the observed data and the results following recommendations are proposed

1. Special care needs to be taken for intellectual disable students in inclusive education setting.
2. The special educator's needs to be properly trained while performing study in special education setting.
3. The ratio of special educator to intellectual disable student to be increased.
4. More time needs to be given for intellectual disable student before performing the test.
5. The time interval of study to be increase so that the students accommodate themselves before commencement of the activity.

6 Conclusion

Before starting the statistical analysis of any data, the data must be checked to see whether they are appropriate for the particular study. In fact, the search of outliers may be considered as a search for the homogeneity of data. Unfortunately, there are no strict statistical rules for definitively identifying outliers. Finding outliers depends on subject area knowledge and an understanding of the data collection process. Thus, it is obvious that if the outlier legitimately removed, they may violate the assumptions of statistical analysis. Thus, the requirement is to include the outlier but in the same time it should not misconstrue the result. There are several methods available to identify the possible outlier. The statistician needs to understand the requirement and employed the proper method to treat the data to arrive the required conclusion.

In the present study the importance of outlier study was explained with a study conducted with the help of special school of BALVIKAS situated in Kodola, Ganjam, INDIA with ten mild special intellectual disable students. The study was conducted with the obtained raw data and sequential approach was performed by using route cause analysis to find the probable cause of variation. Sequentially, outlier study followed by ANOVA study was conducted to arrive a possible conclusion. The one way ANOVA study was conducted with excluding and including the outlier data. The P value was calculated and the significant of P value with respect to outlier was explained. A conclusion was derived and the study was repeated. The study showed a systematic approach and demonstrates the process of handling outlier in data while performing statistical research.

Disclaimer

Acknowledgements

Competing Interests

Authors have declared that no competing interests exist.

References

[1]   Aguinis H, Gottfredson RK, Joo H. Best-practice recommendations for defining, identifying and handling outliers. Organizational Research Methods. 2013;16:270–301.

[2]   Bakker M, Wicherts JM. Outlier removal, sum scores and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. Psychological Methods. 2014;19:409–427.

[3]   Barnett V. The study of outliers: Purpose and model. Journal of the Royal Statistical Society. Series C (Applied Statistics). 1978;27:242–250.

[4]   Barnett V, Lewis T. Outliers in statistical data. Wiley, New York, NY; 1994.

[5]   Hollenbeck JR, DeRue DS, Mannor M. Statistical power and parameter stability when subjects are few and tests are many. The Journal of Applied Psychology. 2006;91:1–5.

[6]   Hitt MA, Harrison JS, Ireland RD, Best A. Attributes of successful and unsuccessful acquisitions of US firms. British Journal of Management. 1998;9:91-114.

[7]   Herman A, Ryan K, Harry J. Best-practice recommendations for defining, identifying and handling outliers. Organizational Research Methods. 2013;1-32.

[8]   Thomas VP, Leander VDM. To remove or not to remove: The impact of outlier handling on significance testing in testosterone data. Adaptive Human Behavior and Physiology. 2017;3:43–60.

[9]   Cornellius YW. Outlier- Why it is important? The tale of the extreme data towards data science; 2018.
(Accessed 20 August 2020)
Available:https://towardsdatascience.com/outlier-why-is-it-important-af58adbefecc

[10]   Anonymous. Engineering statistics hand book. NIST/SEMATECH e-Handbook of Statistical Methods NIST is an agency of the U.S. Department of Commerce; 2012.
(Accessed 20 August 2020)
Available:https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

[11]   Iglewicz B, Hoaglin David. Volume 16: How to detect and handle outliers. The ASQC Basic References in Quality Control: Statistical Techniques. Edward F. Mykytka, Ph.D., Editor; 1993.

[12]   Aguinis H, Pierce CA, Bosco FA, Muslin IS. First decade of Organizational Research Methods: Trends in design, measurement and data-analysis topics. Organizational Research Methods. 2009;12:69-112.

[13] Cohen J, Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd Ed. Mahwah, NJ: Lawrence Erlbaum; 2003.

[14] Frost Jim. Hypothesis testing: An intuitive guide for making data drives decisions. Statistics by Jim Publishing State College, Pennysalvia, U.S.A; 2020.

[15] Bollen KA, Jackman RW. Regression diagnostics: An expository treatment of outliers and influential cases. Modern Methods of Data Analysis. 1990;257-291.

[16] Osborne JW, Overbay A. The power of outliers (and why researchers should ALWAYS check for them). Practical Assessment, Research and Evaluation. 2004;9(6).

[17] Sang KK, Jong HK. Statistical data preparation: Management of missing values and outliers. Korean J Anesthesiol. 2017;70(4):407-411.