Iraqi Journal of Statistical Science (9) 2006
P.P. [58-74]

DETECTION AND TREATMENT OF OUTLIERS IN DATA SETS

Tara Ahmed H. Chawsheen* **Ivan Subhi Latif ****

ABSTRACT

In this paper, we shall try to determine outliers and pinpoint its source of existence by using Box-Whisker plots technique which is an effective approach to detect and treat outliers. Thus the researchers prove that the Box-Whisker-Plot is the most effective method among other methods used in this research which is the hypothesis of the paper.

* *Assistant Lecture-Department of Statistics, College of Administration and Economics, University of Salahaddin.

** Assistant Lecture-Department of Mathematics, College of Education, University of Salahaddin.

Received: 10/ 6 /2005 ____________________Accepted: 30/ 8 / 2005

Iraqi Journal of Statistical Science (9) 2006
P.P. [58-74]

الكشف ومعالجة الشواذ في مجاميع البيانات

الملخص

في هذا البحث نحاول التعرف على الشواذ و تحديد مستوياته و التحري عنه و كيفية التعامل معه في حالة وجوده في البيانات ، ذلك بالاعتماد على طريقة المعرفة بطريقة بوكس–ويسكر–بلوت و التي تعتبر من افضل الطرائق للتحري و التعامل مع الشواذ. و قد توصل الباحثان الى ان هذه الطريقة هي انجح الطرائق المستخدمة في هذا البحث و تنسجم مع البحث وفرضيته.

1- Introduction

Most empirical data bases include a certain amount of exceptional values, generally termed as "outliers." The isolation of outliers is important both for improving the quality of original data and reducing the impact of outlying values in the process of analyzing databases, because many statistical data may include some observations which deviate from the general trend to a certain degree and are referred to as "outliers."

Practically, nearly all experimental data samples are subject to contamination by outliers which theoretically reduce the efficiency, and reliability of statistical methods.

The idea of the existence of outliers was first realized in the mid eighteenth century when Boscovich attempted to determine the elliptical shape of earth. He was able to obtain ten different measurements two of which were discarded for their extreme values, and then calculated the average of the remaining eight data.

Outliers are investigated to see if a reason for their unusual behavior can be found. Sometimes outliers have "bad" values occurring as a result of unusual but explainable events.

Examples include faulty measurement, incorrect recording of data or failure of a measuring instrument. This being the case, the outlier is corrected, or deleted from the data set. [7]

Mark and Kandel (2004)[6], emphasized the importance of outlier frequency in a slightly different definition. High and Robin (2004) show that it is a fact of life that data are not well-behaved. 'Outliers'—unusual data values—pop up in most research projects involve data collection.

This is especially true in observational studies where data may naturally have unusual values, even if they come from reputable sources.[3]

Developing techniques to look for outliers and understanding how the impact data analyses are extremely important parts of a thorough analysis especially when statistical techniques are applied to the data. In this paper, we try to determine outliers and specify the sources of its existence.

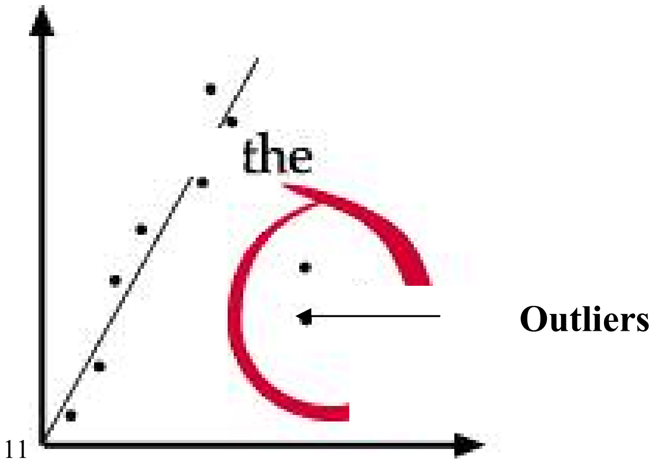
2- Fundamental Concepts

The fundamental concepts that are embodied in this subject can be outlined as below.

2-1 Outliers

An observation in which its standardized residual is large relative to other observations in the data set, it is considered an outlier that lies at a distance from the rest of the data as shown in Figure No. (1).[7]

Figure No. (1)
Examples of outliers



Source: Geres, Fuzzy (2004). "Outliers." Fuzzy Company, Chicago. P.1.(via internet).

2-2 High – Leverage Point

A large leverage that lies away from the center of points in the *X* space *(X Axis)*, is regarded as an outlier in the *X* space.

2-3 Good Leverage Point

The data that usually lies on the linear trend set by the majority of the data is called a good leverage point.

2-4 Bad Leverage Point

If unusual points appear in both *(x_i, y_i)* axis, they are called bad leverage points.

2-5 Influential Observation

A point is defined to be influential if it doesn't conform with the rest of the data. Another way to look at it is: do your results change substantially when computing them with or without such an observation? If so, it is considered an influential observation.

** In general, if the observation diverges significantly in the X- direction it is called a high leverage point and if it diverges in the Y- direction it is known as an influential observation.*

2-6 Extreme Value

It is not certain that those values do not belong to the others because they are either far too big or far too small, Figure No. (2).[8]

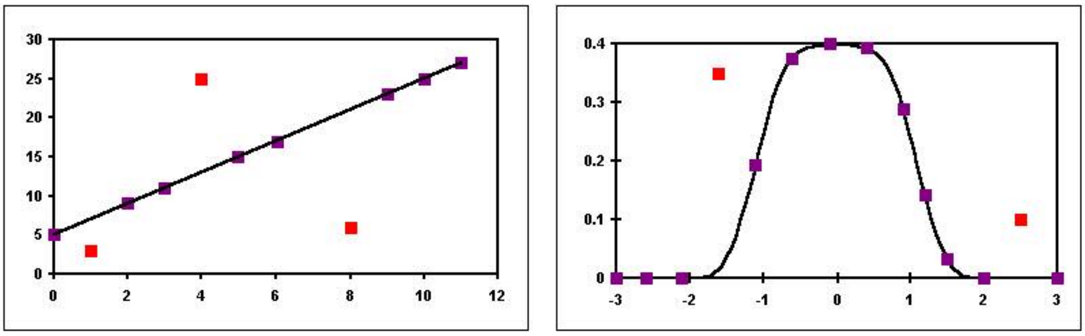


Figure No. (2)

Examples of outliers Note: Outliers are indicated in bold

Source: Fallon, A. and Christian Spada (1997). "Environmental Sampling and Monitoring Primer." ***Gallagher Publication.*** **Vermont.** ***p.2. (via internet)***

3- Source of Outliers

The possible source of outliers can be summarized by recording and measuring errors, incorrect distribution assumption, unknown data structure or just an incidental phenomenon.

Recording and measuring errors are often the first source of suspected outliers.

Incorrect assumption about the data distribution can lead to mislabeling them as outliers. The data which do not fit well into the assumed distribution may fit well into a different set of distributions, see [2][4][8].

Unknown data structure and correlations can cause apparent outliers. A data set could be made up of subsets with two different mechanisms; the first one is that data come from some heavy tailed distribution such as student's t. As such any observation in such a distribution is in no way erroneous. The second mechanism is that data arise from two distributions. One of these, the 'basic distribution', generates 'good' observation, while, the other one 'contaminating distribution', generates 'contaminants.' If the contaminating distribution has tails which are heavier than those of the basic distribution, then there will be a tendency for contaminators to be outliers – that is separable from the good observations, which will then constitute the "inliers". The data set indicated above should be analyzed independently of each other.

4- Outlier Detection

Visual inspection of scatter plots is the most common approach to outlier detection.

Making an analogy between unsupervised and supervised methods of machine learning, two types of detection methods can be distinguished: univariate methods, which examine each variable individually, and multivariate methods, that take into account associations between variables in the same dataset. In both methods a value is considered to be an outlier if it is far away from other values of the same attribute.[1]

The information in a frequency distribution is often graphed more easily. That is if the distribution is graphed the presentation will be more appealing. One very common graphical presentation is the histogram. Tukey [4] introduced a method of organizing interval–scaled data and called it "stem-and-leaf display". This is viewed as an alternative to the Histogram and it is useful in preliminary analysis only, for more information see [1][8].

The Mahalanobis Distance [10] is another method which is used for detecting outliers, and the main reasons for using it is its sensitivity to inter-variable changes in the training data. In addition, it is measured in terms of standard deviations from the mean of the training sample. This method is not perfect though, and in fact there are a number of drawbacks. In this respect, Rouseeuw & Leroy (1987), introduced robust distance to detect outliers since it is very sensitive to outliers.[11]

As mentioned earlier, visual inspection is not a viable method for actual discriminate analysis applications and detecting outliers. After that, many statistical measures are developed in order to determine the magnitude of outliers, Standardized residuals and DIFFTS (different fits), that is why they are so widely recommended.[7], [9]

Visual detection of outliers suffers from two basic limitations subjectiveness and poor scalability. Analysts have to apply their own subjective perception in order to determine the parameters such as the "very far away" and "low frequency." manual inspection of scatter plots for every variable is also an extremely time-consuming task and, not suitable for most commercial databases containing hundreds of numeric and nominal attributes.

An objective and quantitative approach to unsupervised detection of numeric outliers is described in [6]. It is based on the graphical technique of constructing a box plot, box-whisker-plots, which represent the median of all the observations and two hinges, (whisker), or medians of each half of the data set.

Most values are expected in the inter quartile range (IQR) or located between the two hinges. Values lying outside the ±1.5 H or 1.5 (IQR) are called "mild outliers" and values outside the boundaries of 3H or 3(IQR) are termed "extreme outliers." While this method represents a practical alternative to manual inspection of each box plot, it can deal only with continuous variables characterized by unimodel probability distributions. The other limitation is imposed by the ternary classification of all values into "extreme outliers", "mild outliers" and "non-outliers." The classification changes abruptly with moving a value across one of the 1.5 H, or 3H boundaries.

An information-theoretic approach to supervised detection of erroneous data has been developed by Guyon et al. The method requires building a prediction model by using one of data mining techniques (e.g., neural networks or decision trees). See [6].

The patterns having the lowest probability to be predicted correctly by the model are unreliable and should be treated as outliers.

However, this approach ignores the fact that data conformity may also depend on the inherent distribution of database attributes and some subjective, user – dependent factors, see [4] and [6].

5- Interquartile Range

The quartiles can be used to create another measurement of variability, the interquartile range, which is defined as follows:

$$H = Q_3 - Q_1$$

This is used to measure the spread of the middle 50 % of the observations. Large values of this statistic indicate that the first and third quartiles are far apart, indicating a high level of variability.

And these quartiles are used in box-whisker-plot for detecting outliers and depends on five statistics: the minimum, the maximum observations, the first, second, and third quartiles. It also depicts other features of a set of data, as shown in Figure (3).

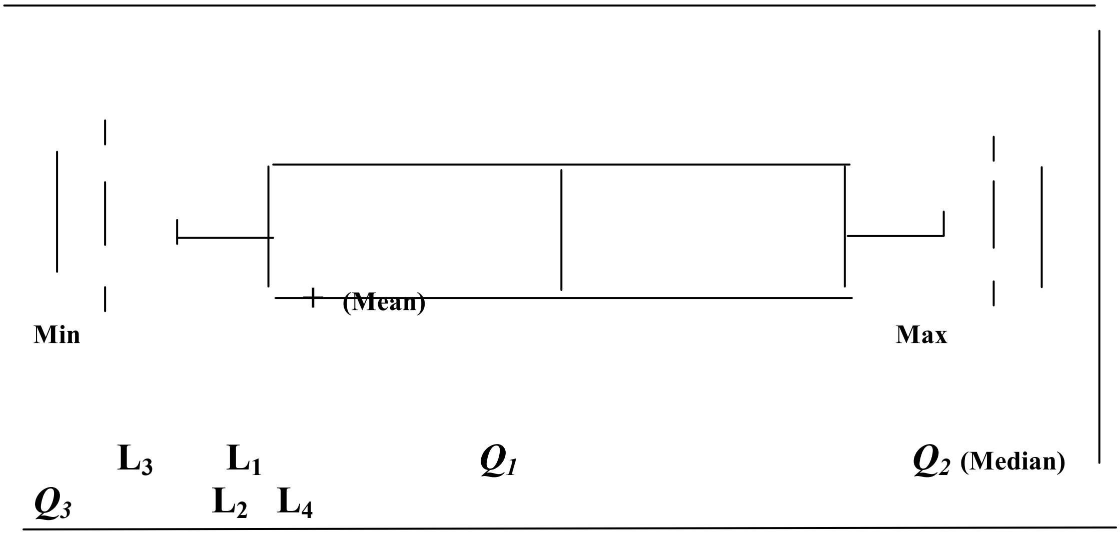


Figure (3)
BOX-WHISKER-PLOT
Source: Originated by the researchers.

The three vertical lines of the box are first, second and third quartiles. The lines extending to the left and right are called whiskers. Any point that lies outside the whiskers is called an outlier.

The whiskers extend outward to the smaller of 1.5 times the interquartile range or to the most extreme point that is not an outlier.

In other words, if any observation lies outside the range of $Q_1 - (1.5\ IQR)$ and $Q_3 + (1.5\ IQR)$ it will be defined as potential outliers and if it lies outside the range $Q_1 - (3\ IQR)$ and $Q_3 + (3\ IQR)$ it is defined as problematic outliers.[5]

6- Treatment of outliers

Effectively working with outliers in numerical data can be a training experience. Various statistical tests have been proposed for detecting and rejecting outliers, because they can cause potential computational and inference problems. A few possible approaches to treat outliers are listed below:-

6-1 Transformation

Transformation data is one way to soften the impact of outliers since the most commonly used expressions, square root and logarithms, modify the larger values to a much greater extent than they do the smaller values.

However, transformations may not fit into the theory of the model or they may affect its interpretation. Taking the log of a variable does more than make a distribution less skewed; it changes the relationship between the original variable and the other variables in the model. In addition, most commonly used transformations require non-negative data or data that is greater than zero, so they do not always provide the answer.

6-2 Deletion

When there are legitimate errors and cannot be corrected, or lie so far outside the range of the data that they distort statistical inferences the outliers should be deleted. When in doubt, we can report model results both with and without outliers to see how much they change.

Data transformation and deletion are important tools, but they should not be viewed as an all-out for distributional problems associated with outliers. Transformations and/or outlier elimination should be an informed choice, not a routine task.

6-3 Accommodation

One very effective plan is to use methods that are robust in the presence of outliers. Non-parametric statistical methods fit into this category and should be more widely applied to continuous or interval data. When outliers are not a problem, simulation studies have indicated their ability to detect significant differences is only slightly smaller than corresponding parametric methods. There are also various forms of robust regression models and computer intensive approaches that deserve further consideration.

7- A Numerical Example

In order to compare the visual inspection of scatter plots and box-whisker-plots for detecting outliers**,** a clinical data sets taken from the Rizgari hospital, Erbil, Iraq, about eye illness in past year (2003). These data are presented in Table (1) below:-

Table (1): The Data Under Study

Y	X_1	X_2	X_3	X_4	X_5
914	62	358	232	160	260
423	109	305	28	**18**	**82**
230	**11**	**143**	87	**14**	119
831	48	246	227	106	149
748	64	348	148	189	254
908	47	322	150	134	138
790	57	344	243	122	250
848	91	385	223	162	314
740	88	375	**576**	105	264
560	96	389	212	164	254
911	80	358	232	118	301
760	**3**	333	226	151	155

Source: Rizgari Hospital Record, Erbil - Iraq, (2003).

A multiple linear regression model which is based on the data in Table (1) above is expressed in the following form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. (7.1),$$

Where

β_j = Unknown parameters (regression coefficients), j = 1, 2, 3, …, 5

Y_i = Blindness and vision weakness, i = 1, 2, 3, …, 12

X_{i1} = Trachoma

X_{i2} = A pinkeye

X_{i3} = Illness of tear gland (body lachrymal)

X_{i4} = Strabismus

X_{i5} = Glaucoma

For detecting outlier in the data under study, the researchers begin with the standardized Residual. According to *staticgraphics* program package, there is one outlier among the entire set of observation; Row (3) belongs to Y_i, X_{i2}, X_{i3} and X_{i4} are as shown in table (2). This Table contains one observation which has a standardized residual value which is greater in absolute value than 2 which measures a number of standard deviations with which each observed value of Y deviates from the fitted model. In this case, there is one standardized residuals greater than 2;[9] thus the value of 3rd row represents an influential point and extreme value simultaneously.

Table (2): Potential Outlier Observations by Using *STATICGRAPHICS* Program Package, Outlier Greater than 2

Row No.	Y	Predicted Y	Residual	Standardized Residual
3	230	382.712	- 152.712	-2.85

Source: Compiled and organized by the researchers.

As for influential points, Table (3) shows four rows: 3, 4, 9, and 12. They are influential points because their values are greater than ׀ DEFFIT ׀ > 2 or their leverage values are greater than three times that of an average data point.

Here, the DFFITE has a limit by which it can tell the existence of influential points. This is called a "cutoff" having a magnitude of 2, but its recommended magnitude is: 2 √(k/2), and if the ׀ DFFITS׀ > cutoff and it is considered as an influential point.

Whereas a leverage is a statistics which measures how influential each observation is in determining the coefficients of the estimated model. In this case, an average data point would have a leverage value equal to 0.5 and there are no data points with more than 3 times the average leverage. Here we find just four data points with unusually large values of DFFITS.

Table (3): Potential Influential Observations by Using *STATICGRAPHICS* Program Package, DFFITS > 2

Row	Y	X_2	X_3	X_4	Leverage	DFFITS
3	230	143	87	14	0.827813	-6.24907
4	831	246	227	106	0.564614	1.91636
9	740	375	576	105	0.905815	- 3.64831
12	760	333	226	151	0.765609	-1.56324

Average leverage of single data Point = 0.5

Source: Compiled and organized by the researchers.

Table No. (4) Shows the extreme values in the stated data.

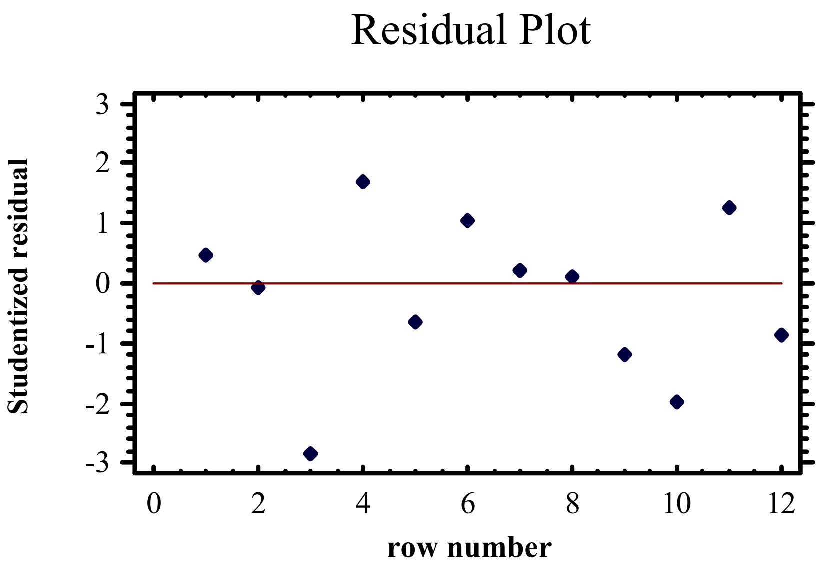
Table (4): Extreme Value, Using *STATICGRAPHICS* Program Package

Row	Y	X_1	X_2	X_3	X_4	X_5
2	0	0	0	0	18	82
3	230	11	0	0	14	0
9	0	0	0	576	0	0
12	0	3	0	0	0	0

Source: Compiled and organized by the researchers.

Figure (4) Confirms the existence of outliers which causes the data to digress the dispersion of the data from normality.

Figure (4): Examples of outliers Note: Outliers are indicated in bold



Source: Originated by the researchers.

With the use of MINITAB program package, the smallest value of Y is equal to 230 and the largest one is 914. The first, second, and third quartiles are 605,775 and 893, respectively. The inter-quartile range is equal to 288. One and one half times the inter quartile range is 1.5(288) = 432. Outliers are defined as any observations that are less than 605-432= 173 and any observations that are larger than 893+432= 1325. The whisker to the left extends to zero, which is the smallest observation that is not an outlier. The whisker to the right extends to 914, which is the largest observation that is not an outlier. Therefore, there are no outliers.

By looking at the Figure No. (5), the data appear to be uniformly distributed. Now for X_{i1}, Figure No. (6) the smallest value is 3 and the largest one is 109, and the first, second and third quartiles are 47.250, 63, and 90.250, respectively. The IQR is equal to 43 and outliers are apparently defined because observations are less than -17.25 and otherwise are larger than 154.75. The whisker to the left extends to 3, which is the smallest observation that is not an outlier. The whisker to the right extends to 109, which is the largest observation for X_{i1}, and it is not an outlier. That means the data are devoid of outlier. Concerning X_{i2}, which is graphed in Figure No. (7), Shows that the smallest value is

143 and the largest one is 389, and the first, second, and third quartiles are 309.250, 346, and 370.750 The IQR is equal to 61.5. Here we notice that the value of third row is less than 217 and it lies outside the whiskers (217 and 432.25) is considered as an outlier.

While Figure No. (8), for X_{i3}, the IQR is equal to 83.5, any observation that lies outside the range of 23.25 and 357.25 is considered to be an outlier. Here we have a value which belongs to row 9 with a magnitude of 576 being greater than 357.25 which is considered an outlier on the right hand side of box-whisker-plot. And it appears to be χ^2 distributed. Now for X_{i4}, its box-whisker-plot shows that there are two outliers see Figure No. (9), the smallest value is equal to 14 and the largest one is equal to 189, while its first quartile is equal to 105.250, second quartile equals 128 and third one is 161.500, IQR is of the magnitude 56.25. If there were any observation that is less than 20.875 and greater than 248.875 it is considered an outlier, since there are two values 18 and 14 belonging to second and third rows of X_{i4} the observations are less than 20.875 so they are considered outliers.

Finally, Figure No. (10) indicates that the data is normally distributed and there is not any outlier, because all observations are less than 446.375 and are greater than -42.625.

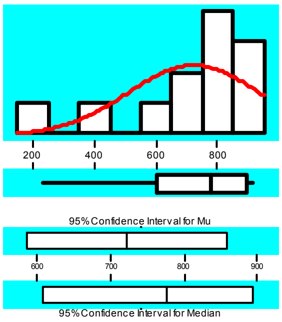


Figure No. (5)

B.V.W.

```
Anderson-Darling Normality Test
   A-Squared:          0.824
   P-Value:            0.023

   Mean              721.917
   StDev             213.140
   Variance          45428.6
   Skewness         -1.41961
   Kurtosis          1.42468
   N                      12

   Minimum           230.000
   1st Quartile      605.000
   Median            775.000
   3rd Quartile      893.000
   Maximum           914.000
95% Confidence Interval for Mu
   586.494           857.339
95% Confidence Interval for Sigma
   150.987           361.886
95% Confidence Interval for Median
   607.361           892.213
```

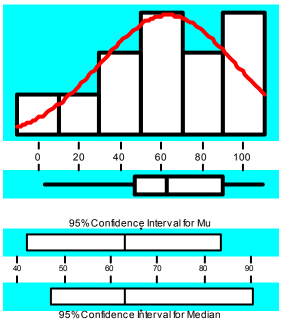


Figure No. (6)

Tra.

```
Anderson-Darling Normality Test
   A-Squared:          0.297
   P-Value:            0.531

   Mean              63.0000
   StDev             32.6497
   Variance             1066
   Skewness         -5.9E-01
   Kurtosis         -2.7E-01
   N                      12

   Minimum             3.000
   1st Quartile       47.250
   Median             63.000
   3rd Quartile       90.250
   Maximum           109.000
95% Confidence Interval for Mu
   42.255             83.745
95% Confidence Interval for Sigma
   23.129             55.435
95% Confidence Interval for Median
   47.263             90.211
```

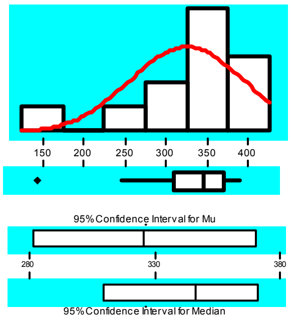


Figure No. (7)

Pin.

```
Anderson-Darling Normality Test
   A-Squared:          0.939
   P-Value:            0.012

   Mean              325.500
   StDev              69.478
   Variance          4827.18
   Skewness         -1.92390
   Kurtosis          3.99253
   N                      12

   Minimum           143.000
   1st Quartile      309.250
   Median            346.000
   3rd Quartile      370.750
   Maximum           389.000
95% Confidence Interval for Mu
   281.356           369.644
95% Confidence Interval for Sigma
   49.218            117.965
95% Confidence Interval for Median
   309.473           370.527
```

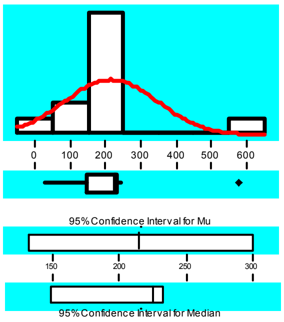


Figure No. (8)

I.T.G.

```
Anderson-Darling Normality Test
   A-Squared:          1.163
   P-Value:            0.003

   Mean              215.333
   StDev             132.145
   Variance          17462.4
   Skewness          1.78120
   Kurtosis          5.50757
   N                      12

   Minimum            28.000
   1st Quartile      148.500
   Median            224.500
   3rd Quartile      232.000
   Maximum           576.000
95% Confidence Interval for Mu
   131.372           299.295
95% Confidence Interval for Sigma
   93.611            224.367
95% Confidence Interval for Median
   148.526           232.000
```

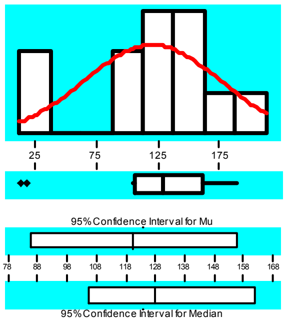


Figure No. (9)

Str.

```
Anderson-Darling Normality Test
  A-Squared:        0.676
  P-Value:          0.057

  Mean            120.250
  StDev            55.009
  Variance        3026.02
  Skewness       -1.11345
  Kurtosis       0.601201
  N                    12

  Minimum          14.000
  1st Quartile    105.250
  Median          128.000
  3rd Quartile    161.500
  Maximum         189.000
 95% Confidence Interval for Mu
  85.299          155.201
95% Confidence Interval for Sigma
  38.968           93.399
95% Confidence Interval for Median
  105.263         161.474
```

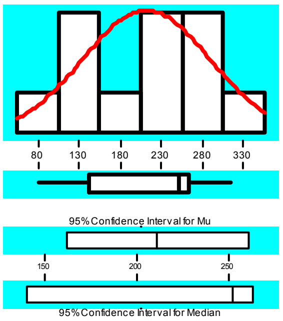


Figure No. (10)

Gla.

```
Anderson-Darling Normality Test
  A-Squared:        0.647
  P-Value:          0.069

  Mean            211.667
  StDev            77.780
  Variance        6049.70
  Skewness       -3.6E-01
  Kurtosis       -1.42202
  N                    12

  Minimum          82.000
  1st Quartile    140.750
  Median          252.000
  3rd Quartile    263.000
  Maximum         314.000
 95% Confidence Interval for Mu
  162.248         261.086
95% Confidence Interval for Sigma
  55.099          132.061
95% Confidence Interval for Median
  140.894         262.948
```

Source: Originated by the researchers.

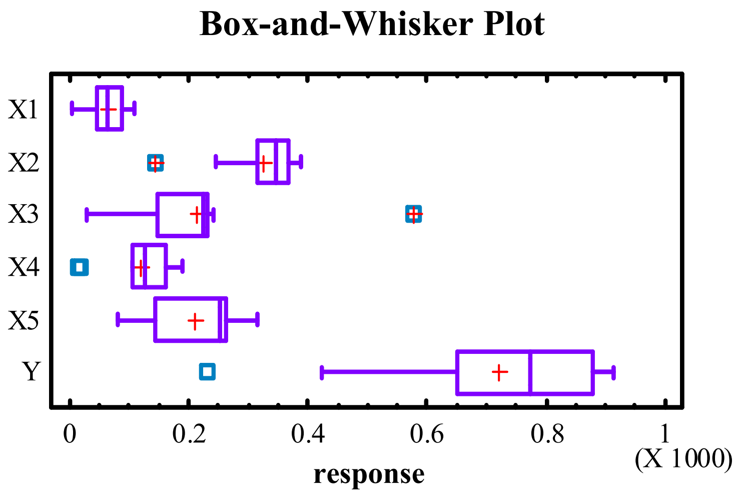
8- Conclusions

Box-Whisker-Plot shows that there are outliers in the data under study (Figure 11), while other measures do not. This explains that it is not necessarily correct that every extreme value to be an influential observation and that not every influential observation is necessarily an outlier.

An influential observation can be recognized in any data if its deletion from the rest of the date changes the result of the analysis. It is also true to say that any point in the data is extreme value if its magnitudes were twice its smallest value.

It is to note that the shape of a box plot is not heavily influenced by a few extreme observations because the median and the other quartiles are not unduly influenced by extreme observations. That is in contrast with the situation in means and variances. Also, outliers affect the real distribution of any data if that data contains influential points.

Thus the researchers prove the research hypothesis.



Source: Originated by the researchers.

Figure (11)

9- Recommendations

In order to enquire further on the merits of the outliers in the future the researchers recommend:

1- The use of Information Criteria to detect outliers.

2- The use of Robust Methods especially Forward Search Algorithm of Least Trimmed Square for finding outliers in any data sets.

10- References

[1] Fallon, A. and C. Spada (1997). "Environmental Sampling and Monitoring Primer." Gallagher publication. Vermont. pp.2-3.

[2] Hawkins, D. M. (1994). Identification of Outliers. Chapman and Hall. New York, p1-2.

[3] High, Robin (2004) "Dealing with 'Outliers' ": How to Maintain Your Data's Integrity.." University of Origon. *Darkwing*. pp.1-2.

[4] Keller, G and B. Warrack (2000). Statistics for Managing and Economics (fifth edition). Pacific Grove. Thomson Learning, Inc., pp.26, 35, 120,123, 669-670.

[5] Keller, G and B. Warrack (2003). Statistics for Managing and Economics (sixth edition). Pacific Grove. Thomson Learning, Inc., pp112, 115, 645-646.

[6] Last, M. and A. Kandel (2004). "Automated detection of outliers in Real-World Data" Ber-Garion University of the Neger and University of South Florida..p2. (via internet).

[7] Montgomery, Douglas C.(1982). Introduction to Linear Regression Analysis. John Wiley and Sons, New York. pp. 70-71.

[8] Siegel, Andrew F. (1997). Practical Business Statistics (3rd edition). Richard D. Irwin, Inc., Homewood, pp43-49.

[9] Statgraphics Plus for Windows (1999). *Statgraphics Users Guide*. Version 4, Microsoft Corp.

[10] Thermo G., "Descriminant Analysis, The Mahlanobis Distance "*Algorithms*, November,2001

[11]القهوتي. وصفي طاهر (2000). بعض طرق التقديرات الحصينة في تحسين الصور الرقمية. رسالة دكتوراه غير منشورة، كلية الادارة والاقتصاد، جامعة بغداد، بغداد. ص ص 18–0.والمعتمد على:

[12] Rousseeue, P. J. and Leroy Annick M. (1987). Robust Regression and Outlier detection. John Wiley and Sons, New York. pp.18-30.