Journal of King Saud University – Science 32 (2020) 2005–2008

Contents lists available at ScienceDirect

Journal of King Saud University – Science

journal homepage: www.sciencedirect.com

Original article

On detecting outliers in complex data using Dixon's test under neutrosophic statistics

Muhammad Aslam

Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah 21551, Saudi Arabia

ARTICLE INFO

Article history:
Received 3 August 2019
Revised 11 December 2019
Accepted 4 February 2020
Available online 14 February 2020

Keywords:
Classical statistics
Neutrosophic statistical interval method
Dixon's test
Testing procedure
A complex system

ABSTRACT

The existing Dixon's test (DT) under classical statistics has been widely applied in a variety of fields. The main target of DT is to recognize the outlier or suspicious observation in the sample. The DT available in the literature is workable when all the observations in the sample or the population are precise, determined and certain. In practice, under the complex system, it may not possible that all observations in the data are determined. In such situations, the existing DT cannot be applied for the detection of the outlier value in the sample. In this paper, we will introduce a new Dixon's test under the neutrosophic statistics is called (NDT). We will present the testing procedure for the proposed test using the neutrosophic statistical interval method. We will discuss its application with the help of an example.

1. Introduction

The estimation and forecasting are done using the data collected from the respective fields of interest. The statistical techniques/methods are applied for the data analysis. The collection of data from the fields is not an easy task. A lot of money, time and effort are needed to collect data from the fields of interest. For example, the collection of the ocean data is not an easy task with the limited resources, see for example, (Jingang et al., 2017). The presence of the potential outliers in the data may mislead the policymakers. In the statistics, an observation that is away from the other observations or has many variations in the sample is known as the outlier. For an accurate analysis of the data, the extreme values should be identified and eliminated from the sample. The elimination of the extreme values of the data is an important field of data mining. According to (Manoj and Senthamarai Kannan, 2013), "The identification of outliers can lead to unexpected knowledge discovery in areas such as credit card fraud detection, criminal behaviors detection, computer intrusion detection, calling card fraud detection, etc. Applications such as outlier detection customized marketing, network intrusion detection, weather prediction; pharmaceutical research and exploration in science databases require the detection of outliers". (Dixon, 1950; Dixon, 1951) proposed the test for the identification of the extreme value in the data. Several authors paid their attention to this issue and presented valuable work in the detection of the outlier using the Dixon's test, see for example, (Verma and Ruiz, 2006; Böhrer, 2008; Manoj and Senthamarai Kannan, 2013; Jabbari Nooghabi, 2017; Jingang et al., 2017; Jabbari Nooghabi, 2019).

In practice, for the variable data, it may not always possible to have the exact observations and attribute data may have the categories. In these situations, data analysis may be overestimated or underestimated. The fuzzy logic, which is the alternative to be applied for the analysis of fuzzy data. Therefore, data mining using fuzzy logic helps to eliminate the outlier from the imprecise data. (Rajeswari et al, 2014) discussed the fuzzy-based approach method to detect the outlier in educational data. More details about the detection of outlier in fuzzy data can be seen in (Hung and Yang, 2006; Gładysz and Kuchta, 2007; Cateni et al., 2007; Yousri et al., 2007; Top et al., 2009; Ramli et al., 2010; D'Errico and Murru, 2012; Kim et al., 2014; Duraj et al., 2019).

The neutrosophic logic which is the generalization of the fuzzy logic is introduced by (Smarandache, 1998). The neutrosophic deals with the measure of the indeterminacy. The neutrosophic statistics, which is the extension of classical statistics, used when there is Neutrosophy in sample or data. The neutrosophic statistics is applied when the observations in the sample or the population are indeterminate, vogues and imprecise. The neutrosophic statis-

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

E-mail address: aslam_ravian@hotmail.com

https://doi.org/10.1016/j.jksus.2020.02.003
1018-3647/

tics is quite effective and adequate to be applied in the uncertainty system, see for example, (Smarandache, 2014; Aslam, 2018b). Chen et al. (2017a) and Chen et al. (2017b) presented the application of the neutrosophic numbers in the area of rock engineering. Aslam (2018b) and Aslam (2019) introduced neutrosophic statistical quality control (NSQC).

As we mentioned earlier, the presence of the outlier can affect the forecasting and estimation of the data. Several authors contributed much work for the detection of the outlier using Dixon's test under classical statistics and using fuzzy logic. By exploring the literature and according to the best of our knowledge, there is no work on Dixon's test under the neutrosophic statistics. In this paper, we will originally propose Dixon's test under the neutrosophic statistics for the identification of the outlier in the data or sample. We will develop Dixon's test under the neutrosophic statistics and explain with the help of an example from the Samsung Company. We expect that the proposed Dixon's test under the neutrosophic statistics will be more effective and adequate to be applied under an uncertainty environment.

2. Preliminaries

Suppose that $X_{nN} = X_n + uI_N;\ X_{nN}\epsilon[X_{nL}, X_{nU}]$ be a neutrosophic random variable having the determinate part X_n and indeterminate part uI_N; $I_N\epsilon[inf, sup]$. Note here that the neutrosophic random variable becomes the random variable under classical statistics when $I_N\epsilon[0, sup]$. We also assume that $X_{nN}\epsilon[X_{nL}, X_{nU}]$ is generated from the neutrosophic normal distribution with mean $\mu_N = \mu + uI_N; \mu_N\epsilon[\mu_L, \mu_U]$ and neutrosophic standard deviation $\sigma_N = \sigma + uI_N; \sigma_N\epsilon[\sigma_L, \sigma_U]$. The neutrosophic mean and standard deviation reduce to mean under classical statistics when $X_{nL} = X_{nU}$. More details about the neutrosophic distributions can be seen in Smarandache (2010) and Aslam (2018a).

3. Dixon's test under NSIM

Let $X_{nN} = X_n + uI_N$ be neutrosophic random variable consists of a determined part X_n and indeterminate part uI_N and $I_N\epsilon[I_L, I_U]$ is an indeterminacy interval. We assumed that $X_{nN}\epsilon[X_{nL}, X_{nU}]$ is selected from the neutrosophic normal distribution. The purpose of the proposed Dixon's test under the neutrosophic statistics is called (NDT) is to detect either the extreme value belongs to the group of observations or not. Like the existing Dixon's test under CS, we have the following assumptions for the proposed NDT

3.1. Assumptions

1. The neutrosophic variable $X_{nN}\epsilon[X_{nL}, X_{nU}]$ is measured from the complex system having fuzzy, interval-based and imprecise values.
2. The neutrosophic variable $X_{nN}\epsilon[X_{nL}, X_{nU}]$ follows the neutrosophic normal distribution.
3. The size of the neutrosophic variable $X_{nN}\epsilon[X_{nL}, X_{nU}]$ should be more than $[3, 3]$.

3.2. Methods

In this section, we discuss the testing procedure of the proposed Dixon's test under NSIM is stated. For the testing, the largest or the smallest sample observations to the outlier, identify that value and put in front. Select a random sample of size $n_N\epsilon[n_L, n_U]$ from the neutrosophic normal distribution and arranged it in ascending order or descending order. Let the neutrosophic sample $X_{nN}\epsilon[X_{nL}, X_{nU}]$ has the values $X_{1N} = X_{1N} + uI_N$, $X_{2N} = X_{2N} + uI_N, \ldots, X_{nN} = X_{nN} + uI_N$. We arrange these neutrosophic observations in ascending order as $X_{1N} \le X_{2N} \le \cdots \le X_{n_N}$, where X_{1N} and X_{n_N} are the smallest and the largest values of the neutrosophic data. The proposed Dixon's test statistic under NSIM is given by

$$R_{10N} = \frac{(X_{n_N} - X_{n_N-1})}{(X_{n_N} - X_{1N})};\ R_{10N}\epsilon[R_{10L}, R_{10U}], X_{n_N}\epsilon[X_{nL}, X_{nU}] \tag{1}$$

$$R_{11N} = \frac{(X_{n_N} - X_{n_N-1})}{(X_{nN} - X_{2N})};\ R_{11N}\epsilon[R_{11L}, R_{11U}], X_{n_N}\epsilon[X_{nL}, X_{nU}] \tag{2}$$

$$R_{21N} = \frac{(X_{n_N} - X_{n_N-2})}{(X_{n_N} - X_{2N})};\ R_{21N}\epsilon[R_{21L}, R_{21U}], X_{n_N}\epsilon[X_{nL}, X_{nU}] \tag{3}$$

$$R_{22N} = \frac{(X_{n_N} - X_{n_N-2})}{(X_{n_N} - X_{3N})};\ R_{22N}\epsilon[R_{22L}, R_{22U}], X_{n_N}\epsilon[X_{nL}, X_{nU}] \tag{4}$$

By following (Dixon, 1950; Dixon, 1951) suggestions, we classify the proposed test statistic with respect to the suitable values of $n_N\epsilon[n_L, n_U]$ as follows

$$R_N^{(n_N)}\epsilon \begin{bmatrix} R_{10N} = \frac{(X_{n_N} - X_{n_N-1})}{(X_{n_N} - X_{1N})}; [3,3] < n_N < [7,7] \\ R_{11N} = \frac{(X_{n_N} - X_{n_N-1})}{(X_{nN} - X_{2N})}; [8,8] < n_N < [10,10] \\ R_{21N} = \frac{(X_{n_N} - X_{n_N-2})}{(X_{n_N} - X_{2N})}; [11,11] < n_N < [13,13] \\ R_{22N} = \frac{(X_{n_N} - X_{n_N-2})}{(X_{n_N} - X_{3N})}; [14,14] < n_N < [30,30] \end{bmatrix};\ R_N^{(n_N)}\epsilon\left[R_L^{(n_N)}, R_U^{(n_N)}\right] \tag{5}$$

The neutrosophic form of the proposed test can be written as

$$R_N^{(n_N)} = A_N + B_N I_N; I_N\epsilon[I_L, I_U] \tag{6}$$

where A_N and $B_N I_N$ are the determinate and indeterminate parts of the proposed test. Note here that the proposed Dixon's test statistic under the neutrosophic statistics is the generalization of the (Dixon, 1950; Dixon, 1951) test statistic under classical statistics. The proposed Dixon's test statistic reduces to (Dixon, 1950; Dixon, 1951) test statistic if $X_{nL} = X_{nU}$ or $I_L = 0$. More details on the existing statistics can be seen in (Kanji, 2006). To test whether the suspected observation is an outlier or not, the neutrosophic$R_N^{(n_N)}$, which is based on the ascending order of observations, $X_{n_N}\epsilon[X_{nL}, X_{nU}]$is computed. According to the test, an observation is declared to be an outlier if the calculated value of $R_N^{(n_N)}$ exceeds the critical value at a specified level of significance α.

4. Example

In this section, we discuss the application of the proposed Dixon's test statistic in the area of neutrosophic statistical quality control (NSQC). The Samsung Company is manufacturing several electronic equipment and mobile phones. This company is very popular in Saudi Arabia due to the high reliability of their product. The Samsung Company is interested to apply the proposed Dixon's test statistic on cell phone batteries failure time. For the inspection of the batteries product, the experimenters at the company have selected a sample at regular interval during the manufacturing process. Due to the complexity of the lifetime measuring process, on several processes, the measurements are not accurate, precise and sometimes in an interval. Therefore, there is Neutrosophy in the failure time of the batteries product. Therefore, for the testing of the outlier in the failure time, the existing Dixon's test statistic under classical statistics cannot be applied. The failure time of 12 batteries having neutrosophic observations is shown in Table 1. The company is expected to have failed time greater than

Table 1
The neutrosophic failure time.

[973,975]	[1214,1220]	[1668,1668]	[1903,1903]
[1056,1056]	[1449,1451]	[1798,1800]	[1921,1925]
[1207,1210]	[1566,1570]	[1867,1870]	[1946,1950]

1000 h. From the data, they note that one failure time is from 973 to 975. The company is interested to test it either it is an outlier or not in the sample. Therefore, we apply the proposed Dixon's test statistic to test the null hypothesis that $X_{1N}\epsilon[973,975]$ is an outlier Vs the alternative hypothesis that it is not an outlier and belong to the same group. There are 12 observations in the sample, therefore, we will use $R_N^{(n_N)}\epsilon\left[R_L^{(n_N)},R_U^{(n_N)}\right]$ based on $R_{21N}\epsilon[R_{21L},R_{21U}]$. The null hypothesis will be rejected if the value of $R_N^{(n_N)}\epsilon\left[R_L^{(n_N)},R_U^{(n_N)}\right]$ based on $R_{21N}\epsilon[R_{21L},R_{21U}]$ will be greater than critical. Some necessary calculations for this data are shown as follows.

Step-1: Use the neutrosophic Kolmogrov-Smirnov (NK-S) test to see either the available data follows the neutrosophic normal distribution or not. The neutrosophic P-value of $X_{nN}\epsilon[X_{nL},X_{nU}]$ is [0.79, 0.80] which indicates that given lifetime data follows the neutrosophic normal distribution.

Step-2: arrange the batteries failure time $n_N \in$ [12, 12] in ascending order, say $[973,975]\leq[1056,1056]\leq\cdots\leq[1946,1950]$ as in Table 1. Let$X_{12N}\epsilon[973,975]$, $X_{10N}\epsilon[1207,1210]$ and $X_{2N}\epsilon[1921,1925]$.

Step-3: Compute $R_N^{(n_N)}=\frac{(X_{12_N}-X_{10N})}{(X_{12_N}-X_{2N})}=\frac{[973,975]-[1207,1210]}{[973,975]-[1921,1925]}=[0.2468,0.2473]$.

Step-4: From (Kanji, 2006), the tabulated value is [0.546, 0.546] when α = 0.05 and $n_N \in$ [12, 12].

Step-5: The computed $R_N^{(n_N)}\epsilon[0.2468,0.2473]$ is smaller than the critical value [0.546, 0.546]. Therefore, it is concluded that the suspected value in the sample is not an outlier.

5. Discussion

We presented the testing procedure of the proposed Dixon's test in Section 4. We note from the calculations of the real data under the neutrosophic statistics; the proposed testing procedure provides the analysis of the data in the indeterminacy interval rather than the exact or determined values as in classical statistics. Chen et al. (2017a) and Chen et al. (2017b) suggested that a method which provides the values in the indeterminacy interval under uncertainty is said to be a more effective and adequate method. The neutrosophic form of the proposed test for the data is: $R_N^{(n_N)}=0.2468+0.2473I_N; I_N\epsilon[0,0.0020]$. Note here that there is measure of indeterminacy 0.0020, which is associated with the proposed test under uncertainty. The analysis of the lifetime of the data suggested that the values of the proposed statistic can be from 0.2468 and 0.2473. On the other hand, the existing testing procedure given by Dixon (1950) and Dixon (1951) which is the special case of the proposed testing procedure provide the exact value of the statistic which is 0.2468. Furthermore, the existing test does not provide the measure of indeterminacy associated with the test. From this comparison, it is clear that the proposed testing procedure is more adequate and effective to be used in uncertainty environment than the testing procedure given by Dixon (1950) and Dixon (1951) under classical statistics.

6. Concluding remarks

In this paper, we introduced a new Dixon's test under the neutrosophic statistics. We presented the testing procedure for the proposed test using the neutrosophic statistical interval method. The proposed testing procedure for the detection of the outlier in the data is more adequate and effective than the testing procedure available in the literature. The proposed testing procedure is more information to be applied when observations in the sample or the data are imprecise and fuzzy. The comparison between the proposed testing procedure and existing testing procedure support claim and coincides with the theory of Chen et al. (2017a) and Chen et al. (2017b). The proposed Dixon's test under the neutrosophic statistics is an effective alternative of the existing Dixon's test under classical statistics. The proposed Dixon's test statistic can be used to detect the outlier in the data for the better inferences from it. We will introduce the proposed Dixon's test in the area of control chart theory as future research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author is deeply thankful to the editor and the reviewers for their valuable suggestions to improve the quality of this manuscript. This work was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant No. (130-244-D1440). The author, therefore, gratefully acknowledge the DSR technical and financial support.

References

Aslam, M., 2018a. Design of sampling plan for exponential distribution under neutrosophic statistical interval method. IEEE Access 6, 64153–64158.
Aslam, M., 2018b. A new sampling plan using neutrosophic process loss consideration. Symmetry 10, 132.
Aslam, M., 2019. Attribute control chart using the repetitive sampling under neutrosophic system. IEEE Access 7, 15367–15374.
BöHRER, A., 2008. One-sided and two-sided critical values for Dixon's outlier test for sample sizes up to n = 30. Econ. Quality Control 23, 5–13.
S. Cateni, V. Colla, M. Vannucci, A fuzzy logic-based method for outliers detection. Artificial Intelligence and Applications, 2007. 605-610.
Chen, J., Ye, J., Du, S., 2017a. Scale effect and anisotropy analyzed for neutrosophic numbers of rock joint roughness coefficient based on neutrosophic statistics. Symmetry 9, 208.
Chen, J., Ye, J., Du, S., Yong, R., 2017b. Expressions of rock joint roughness coefficient using neutrosophic interval statistical numbers. Symmetry 9, 123.
D'errico, G.E., Murru, N., 2012. Fuzzy treatment of candidate outliers in measurements. Advan. Fuzzy Syst. https://doi.org/10.1155/2012/783843.
Dixon, W.J., 1950. Analysis of extreme values. Ann. Mathematical Statistics 21, 488–506.
Dixon, W.J., 1951. Ratios involving extreme values. Ann. Mathematical Statistics 22, 68–78.
Duraj, A., Niewiadomski, A., Szczepaniak, P.S., 2019. Detection of outlier information by the use of linguistic summaries based on classic and interval-valued fuzzy sets. Int. J. Intelligent Systems 34, 415–438.
B. Gładysz, D. Kuchta, Outliers detection in selected fuzzy regression models. International Workshop on Fuzzy Logic and Applications, 2007. Springer, 211-218.
Hung, W.-L., Yang, M.-S., 2006. An omission approach for detecting outliers in fuzzy regression models. Fuzzy Sets and Systems 157, 3109–3122.
Jabbari Nooghabi, M., 2017. Detecting outliers in exponentiated Pareto distribution. J. Sci., Islamic Republic of Iran 28, 267–272.
M. Jabbari Nooghabi, On detecting outliers in the Pareto distribution. J. Statistical Comput. Simul., 2019, 1–16.
Jingang, J., Lu, S., Zhongya, F., Jiaguo, Q., 2017. Outlier detection and sequence reconstruction in continuous time series of ocean observation data based on difference analysis and the Dixon criterion. Limnol. Oceanography: Methods 15, 916–927.
G.K. Kanji, 2006. 100 statistical tests, Sage.
Y.K. Kim, S.Y. Lee, S. Seo, K.M. Lee, Fuzzy logic-based outlier detection for biomedical data. 2014 International Conference on Fuzzy Theory and Its Applications (iFUZZY2014), 2014. IEEE, 117-121.
Manoj, K., Senthamarai Kannan, K., 2013. Comparison of methods for detecting outliers. Int. J. Scientific Eng. Res. 4, 709–714.

A. Rajeswari, M. Sridevi, C. Deisy, Outliers detection on educational data using fuzzy association rule mining. Proceedings of International Conference on Advanced in Computer Communication and Information Science (ACCIS-14), 2014. 1-9.

Ramli, N., Mohamad, D., Sulaiman, N.H., 2010. Evaluation of teaching performance with outliers data using fuzzy approach. Procedia-Social Behav. Sci. 8, 190–197.

F. Smarandache, Neutrosophy. Neutrosophic Probability, Set, and Logic, ProQuest Information & Learning. Ann Arbor, Michigan, USA, 105, 1998. 118–123.

F. Smarandache, Neutrosophic Logic-A Generalization of the Intuitionistic Fuzzy Logic. Multispace & Multistructure. Neutrosophic Transdisciplinarity (100 Collected Papers of Science), 2010, 4, 396.

F. Smarandache, Introduction to neutrosophic statistics, Infinite Study, 2014.

M.K. Top, Y. Trouiller, V. Farys, D. Fuard, E. Yesilada, C. Martinelli, M. Said, F. Foussadie,r P. Schiavone, Outliers detection by fuzzy classification method for model building. Metrology, Inspection, and Process Control for Microlithography XXIII, 2009. International Society for Optics and Photonics, 72721G.

Verma, S.P., Ruiz, A.Q., 2006. Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. Revista Mexicana de Ciencias Geológicas 23, 133–161.

N.A. Yousri, M.A. Ismail, M.S. Kamel, Fuzzy outlier analysis a combined clustering-outlier detection approach. 2007 IEEE international conference on systems, man and cybernetics, 2007. IEEE, 412-418.