THE MEAN METHOD: A SPECIFIC OUTLIER BOUNDARY FOR ARBITRARY DISTRIBUTIONS

Louis M. Houston, Karl H. Hasenstein and Naresh T. Deoli
The University of Louisiana at Lafayette, U.S.A.
The Louisiana Accelerator Center
Email: houston@louisiana.edu

ABSTRACT

We prove theorems that delineate the upper and lower outlier boundaries for an arbitrary distribution of real numbers as twice the means of the positive and negative sets derived from the distribution, respectively. The essence of the work is based on a theorem on categories. The derived method of detecting outlier boundaries is independent of data sampling.

Keywords: outlier, category, sigma, interquartile, Chebyshev's Inequality, median

1. INTRODUCTION

An outlier is a data point that is dissimilar to the overall pattern of the data and it is a point that is distinct from the representative observations [1]. When an outlier occurs within a distribution, it may be due to experimental error or to the presence of a combination of different statistical sources [2], [3]. A convenient definition of an outlier is a point which falls more than 1.5 times the interquartile range above the third quartile or below the first quartile [4], [5]. Another popular definition is based on Chebyshev's Inequality that guarantees that for any distribution, no more than $1/k^2$ of the distribution's values can be more than k standard deviations away from the mean [6], [7].

Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution [8]. In the former case one wishes to discard them or use statistics that are robust to outliers [e.g. the median], while in the latter case they indicate that the distribution has high kurtosis and that one should be very cautious in using tools or intuitions that assume a normal distribution [9]. A frequent cause of outliers is a mixture of two distribution, which may be two distinct sub-populations, or may indicate 'correct trial' versus 'measurement error'; this is modeled by a mixture model [10], [11].

In most larger samplings of data, some data points will be farther away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distributions, or it may be that some observations are far from the center of the data [12]. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected (and not due to any anomalous condition) [13].

Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low.

However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

Estimators capable of coping with outliers are said to be robust: the median is a robust statistic, while the mean is not. We prove in this paper that the susceptibility of the mean to outliers is the key to defining the outlier [14], [15].

Methods for detecting outliers either have a complex implementation, like the modified Thompson Tau test or Principal Component Analysis or they rely on interpretation [16], [17], like the interquartile method or they depend upon the sampling of the data set, like the Chebyshev method. The method presented in this paper that we refer to as the mean method is simple to implement, objective and independent of the sampling of the data set. The derivation of the method is purely mathematical and rigorous.

A distribution has frequency and spacing. By extracting the set from the distribution, we remove frequency as a variable. That leaves spacing, which we assert to be the determining factor of an outlier. The essence of the derivation is a novel theorem on categories. Using this theorem, we are able to categorize an outlier boundary for one data spacing and extend that categorization to other spacing's. We point out that this theorem has no formal connection to Category Theory [18].

2. OUTLIER MATHEMATICS

The Necessary Theorems:

THEOREM I. Given a continuous function $f(x)$ with a real domain, if $\exists$ a category C that is valid for $f(x')$, then it is valid for $f(x)\ \ \forall x$.

Proof. Because $f(x)$ is continuous, we must have $C : f(x') = C : f(x' + \in)$, where $\in$ is an infinitesimal real number. That is, there is no basis to categorize $f(x')$ differently from $f(x' + \in)$. This can be repeated successively to yield: $C : f(x') = C : f(x' + \in + \in + ...)$. □

THEOREM II. Given a continuous operator $O(\vec{x})$ with a real vector domain, if $\exists$ a category C that is valid for $O(\vec{x}')$, then it is valid for $O(\vec{x})\ \ \forall\vec{x}$.

Proof. Because $O(\vec{x})$ is continuous, we must have $C:O(\vec{x}')=C:O(\vec{x}'+\vec{\in})$, where $\vec{\in}$ is an infinitesimal real vector. That is, there is no basis to categorize $O(\vec{x}')$ differently from $O(\vec{x}'+\vec{\in})$. This can be repeated successively to yield: $C:O(\vec{x}')=C:O(\vec{x}'+\vec{\in}+\vec{\in}+...)$. □

AXIOM I. A distribution has no outliers if the data are uniformly spaced.

THEOREM III. If the data are uniformly spaced, then the outlier boundaries must be near the extremes of the data.

Proof. Axiom I. □

Theorems I. and II. are analogous for functions and operators. We can demonstrate applications of Theorem I. for the following cases.

(case 1):

$$f(x)=i^{ix},\ x\in\mathbb{R}. \tag{1}$$

Since $C:f(0)\in\mathbb{R}$, by Theorem I., we must have $C:f(x)\in\mathbb{R}\ \ \forall x$. We can confirm this by evaluating $f(x)$. Euler's formula is

$$e^{iy}=\cos y+i\sin y. \tag{2}$$

Therefore,

$$e^{i\pi/2}=i. \tag{3}$$

$$f(x)=i^{ix}=\left(e^{i\pi/2}\right)^{ix}=e^{-(\pi/2)x}. \tag{4}$$

(case 2):

$$f(x)=-1^{\lfloor x\rfloor},\ x\in\mathbb{R}^{+}. \tag{5}$$

The function $f(x)$ is not continuous, so Theorem I. does not apply. For example, $C:f(0)>0$ does not imply $C:f(x)>0,\forall x$.

(case 3):

$$f(x) = -1^{2\lfloor x \rfloor},\ x \in \mathbb{R}^+. \tag{6}$$

The function $f(x)$ is continuous, so Theorem I. does apply. For example, $C: f(0) > 0$ does imply $C: f(x) > 0, \forall x$.

The Derivation of the Upper Outlier Boundary:

Given a distribution of data D consisting of positive real numbers, derive the set $\{y_i\}$ by removing redundancy from D. Consider the following:

$$D \to \{y_i\},\ y_i \in \mathbb{R}^+. \tag{7}$$

$$\{y_i\} = \{x_i\} \cup \{z_0\},\ z_0 > x_{\max}. \tag{8}$$

$\exists\ m \in \mathbb{Z}^+$ such that $10^m y_i \in \mathbb{Z}^+$. Including set ordering, we write $\vec{x} \Leftrightarrow \{x_i\}$ and $\vec{y} \Leftrightarrow \{y_i\}$. Let $2\langle y_i \rangle$ represent twice the mean of $\{y_i\}$. Then we can write:

$$2\langle 10^m y_i \rangle = O(\vec{x}) = \frac{2\left(\sum_{i=1}^{N} 10^m x_i + 10^m z_0\right)}{N+1}. \tag{9}$$

Consider equation (9) for a one-element vector that we denote with $\vec{x}_1$:

$$O(\vec{x}_1) = 10^m (x_1 + z_0) \geq 1 + 10^m z_0. \tag{10}$$

The distribution that generates the two elements $\{10^m x_1\} \cup \{10^m z_0\}$ must have uniform spacing. Therefore, by Theorem III. and equation (10), we can categorize $O(\vec{x}_1)$ as an upper outlier boundary. Since $O(\vec{x})$ is continuous and $\vec{x}$ is a real vector, by Theorem II., we can categorize $O(\vec{x})$ as an upper outlier boundary for all $\vec{x}$. From this result, it follows that $2\langle y_i \rangle$ is an upper outlier boundary for all $\vec{x}$.

The Outlier Boundaries for Real Numbers:

Given a distribution D of real numbers, derive the set $\{w_i\}$ from the distribution. We have $\{w_i^+\} \in \mathbb{R}^+$ and $\{w_i^-\} \in \mathbb{R}^-$. Define the operation $|v|$ as the magnitude of v if v is a

number and the cardinality of v if v is a set. Then, $2\langle|w_i^+|\rangle$ and $-2\langle|w_i^-|\rangle$ are the upper and lower outlier boundaries, respectively.

The Outlier Boundary for a Symmetric Distribution:

THEOREM IV. Given $\{x_i\} = \{x_i^+\} \cup \{x_i^-\}$, $x_i \in \mathbb{R}$, if $\{x_i\}$ is symmetric, then $\langle|x_i^+|\rangle = \langle|x_i^-|\rangle = \langle|x_i|\rangle$.

Proof. $\langle|x_i|\rangle = \dfrac{\sum_{i=1}^{2N}|x_i|}{2N} = \dfrac{\sum_{i=1}^{N}\left(|x_i^+| + |x_i^-|\right)}{2N} = \dfrac{1}{2}\left(\langle|x_i^+|\rangle + \langle|x_i^-|\rangle\right)$.

$\{x_i\}$ symmetric implies $\langle|x_i^+|\rangle = \langle|x_i^-|\rangle = a$. $\therefore \langle|x_i|\rangle = \dfrac{1}{2}(2a) = a$. □

A Graph of Outlier Behavior:

Using equation (9), we can derive the relation:

$$O(\vec{x}) = \frac{2\left(\sum_{i=1}^{N} 10^m x_i + 10^m z_0\right)}{N+1} \geq \frac{2\left(\sum_{i=1}^{N} i + 10^m z_0\right)}{N+1}. \tag{11}$$

Letting $z_0' = 10^m z_0$, we find:

$$O(\vec{x}) \geq N + \frac{2z_0'}{N+1}. \tag{12}$$

This provides a model of the outlier boundary as a function of N. If we write:

$$g(u) = u + \frac{2z_0'}{u+1}, N \to u, \tag{13}$$

we can examine $g(u), u \in \mathbb{R}^+$, $u_{\max} < z_0'$ graphically as an interpolation of the lower bound of $O(\vec{x})$.

This is shown in Figure 1 for $z_0' = 100$. Observe that the function is asymmetric and the outlier boundary is minimal for $u \approx 30$. The outlier boundary is larger as the sample size increases above $u \approx 30$. Also observe that the outlier boundary is near $u_{\max}$ for the extremes with uniform spacing, consistent with Theorem III.

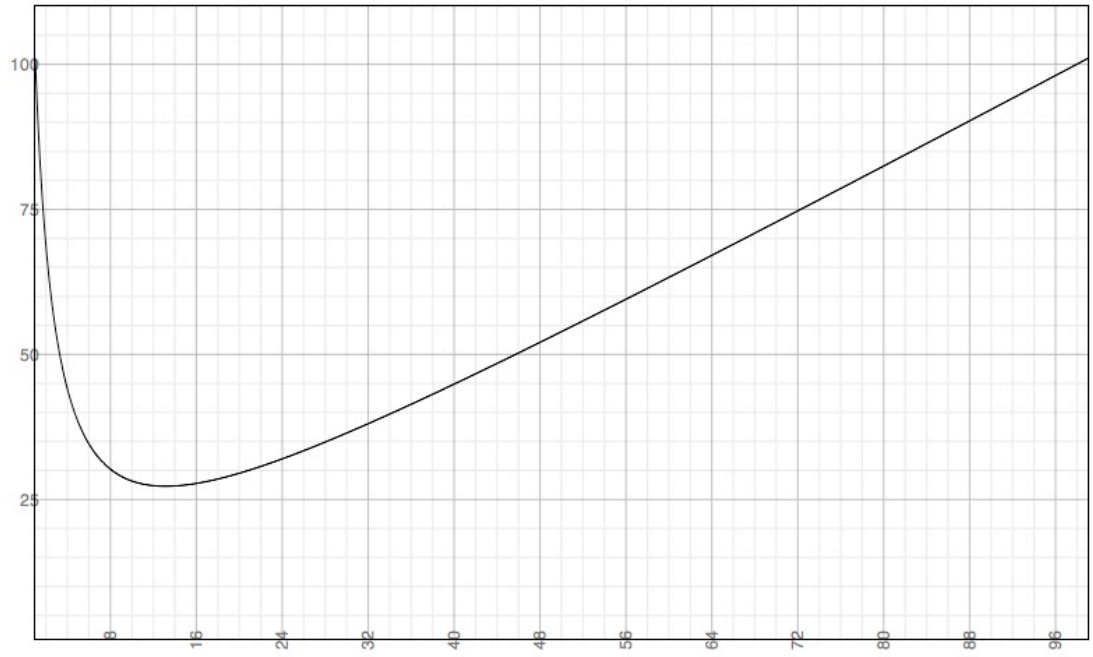

Fig. 1. A graph of $g(u) = u + \frac{2z_0'}{u+1}, N \to u$, $u_{\max} < z_0'$ for $z_0' = 100$.

3. APPLICATIONS

(case1): synthetic data

(a) $D = \{-0.3,-0.3,-0.5,-0.5,0.5,0.8,0.8,10\}$, $D \to \{w_i\}$. (14)

$\{w_i^+\} = \{0.3,0.5,0.8,10\}$. (15)

$\{w_i^-\} = \{-0.3,-0.5\}$. (16)

The mean of D is

$\mu = 1.3125$. (17)

We can estimate the standard deviation σ by using the equation:

$$\sigma \approx \sqrt{\frac{\sum_{i=1}^{N}(D_i - \mu)^2}{N-1}} = 3.55 . \quad (18)$$

$\mu + 2\sigma = 8.41$. (19)

So $\mu + 2\sigma$ is greater than 88% of the data.

Using the 'mean' method presented in this paper, we have

$$\left\langle \left| w_i^+ \right| \right\rangle = \frac{0.5 + 0.8 + 10}{3} = 3.77 . \tag{20}$$

or the upper outlier boundary is:

$$2\left\langle \left| w_i^+ \right| \right\rangle = 7.54 . \tag{21}$$

Similarly, we have

$$\left\langle \left| w_i^- \right| \right\rangle = \frac{0.3 + 0.5}{2} = 0.4 . \tag{22}$$

or the lower outlier boundary is:

$$-2\left\langle \left| w_i^- \right| \right\rangle = -0.8 . \tag{23}$$

Consequently, based on the Chebyshev method, that places the outlier boundary at $\mu + 2\sigma$ for small data sets and $\mu + 3\sigma$ for larger data sets, there are no outliers. Based on the 'mean' method, 10 is the outlier. Observe that $2\left\langle \left| w_i^+ \right| \right\rangle$ is greater than 75% of the positive data in D.

(b) $D = \{500\}, \quad D \to \{w_i\}$. (24)

$$\{w_i^+\} = \{500\} . \tag{25}$$

$$\{w_i^-\} = \{\varnothing\} . \tag{26}$$

The mean of D is

$$\mu = 500 . \tag{27}$$

The standard deviation cannot be estimated because there is only on sample. Therefore, we cannot use the Chebyshev method to detect an outlier boundary. However, with the mean method, we find the upper outlier boundary to be

$$2\left\langle \left| w_i^+ \right| \right\rangle = 1000 . \tag{28}$$

The lower outlier boundary is

$$-2\left\langle \left| w_i^- \right| \right\rangle = 0 . \tag{29}$$

(case 2): real data-a set of sizes of arbitrary email messages in kilobytes

$D = \{y_i\} = \{5,165,84,536,14,3,464,37,11,89,2,12,19,18,17,76,7,15,56,16,67,4,28,26,6,$

$29,86,30,181,46,8,10,169,35,2000,51,405\}$

$N = 37,\ \mu = 130.46.$ (30)

We can use equation (18) to estimate σ:

$\sigma = 340.$ (31)

$\mu + 2\sigma = 810.46.$ (32)

$\mu + 3\sigma = 1150.46.$ (33)

Both $\mu + 2\sigma$ and $\mu + 3\sigma$ are greater than 97% of the data. Using the mean method, the upper outlier boundary is:

$2\langle y_i \rangle = 260.9.$ (34)

Consequently, based on the Chebyshev method, 2000 is an outlier. Based on the mean method, the outliers are $\{405,464,536,2000\}$. The percentage of the data that is less than $2\langle y_i \rangle$ is 89%.

Comments on Applications

For a normal distribution, the outlier boundaries predicted by the Chebyshev method are 75% of the data for less than 22 samples and 89% of the data for greater than 22 samples [19]. This corresponded to the results of the mean method for the arbitrary distributions presented as example applications. However, we were unable to measure a σ that yielded reasonable outlier boundaries based on the Chebyshev method. In the case of one sample point, the Chebyshev method cannot be applied because there is no way to estimate σ. However, in this case, the mean method gives definitive results.

4. CONCLUSION

We base our derivation of a specific outlier boundary on the assertion that an outlier depends on the spacing and not the frequency of the data. Using this Axiom, we prove theorems that define an outlier boundary as twice the mean of the set derived from a distribution. This result is simple to implement, but based on rigorous mathematics. We

use the lack of robustness of the mean to detect an outlier in contrast to methods that seek to avoid outliers by using robust estimators like the median. The essential result is based on a powerful theorem on categories that is not formally connected to Category Theory. We also present examples that indicate the broad applicability of the category theorem. In this paper, we present various examples of the application of the outlier boundary detection method (i.e. the mean method) and get results that are consistent with a visual interpretation of the data and the results of the two sigma and three sigma statistics for normal distributions based on the Chebyshev Inequality. We demonstrate the effectiveness of the method on data from which a good estimate of sigma is not derivable, so that outlier detection methods that rely on sigma are unsuccessful. This is important because there are many situations in which data samples may be too small or irregularly sampled to get a proper estimate of sigma, but outliers are still present.

REFERENCES

[1] Moore, D.S. and McCabe, G.P. *Introduction to the Practice of Statistics*, 3rd ed. New York: W.H. Freeman, 1999.

[2] Pierce, B., "Criterion for the Rejection of Doubtful Observations", Astronomical Journal II **45**(1852).

[3] Pierce, B., (May 1877-May 1878). "On Pierce's Criterion". Proceedings of the American Academy of Arts and Sciences **13**: 348-351.

[4] Dawson, R. How Significant is a Boxplot Outlier?, Journal of Statistics Education, Volume **19**, Number 2(2011).

[5] Hoaglin, D., Tukey, J. W. Performance of some resistant rules for outlier labeling. Journal of the American Statistical Association, Vol. **81**, No. 396 (Dec., 1986), 991-999.

[6] Isii K. (1959) On a method for generalizations of Tchebcheff's inequality. Ann Inst Stat Math **10**: 65-88.

[7] Ferentinos K. (1982) "On Tchebycheff type inequalities". Trabojos Estadist Investigacion Oper, **33**:125-132.

[8] Grubbs, F.E. (February 1969), "Procedures for detecting outlying observations in samples", Technometrics **11** (1): 1-21.

[9] DeCarlo, L.T., On the Meaning and Use of Kurtosis, Psychological Methods, 1997, Vol. **2**, No. 3, 292-307.

[10] Cerioli, A., Multivariate Outlier Detection With High-Breakdown Estimators, Journal of the American Statistical Association, 2010, vol. **105**, issue 489, pages 147-156.

[11] Schwager, S.J. and Margolin, G.H. (1982). Detection of multivariate normal outliers. Annals of Statistics, **10**, 943-954.

[12] Acuna, E., Rodriguez, C. A Meta analysis study of outlier detection methods in classification. Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, 2004.

[13] Knorr, E.M.; Ng, R.T.; Tucakov, V. (2000). "Distance-based outliers: Algorithms and Applications". The VLDB Journal the International Journal on Very Large Data Bases **8**(3-4): 237.

[14] Brys, G., Hubert, M., Rousseeuw, P.J. A robustification of independent component analysis. Journal of Chemometrics 2005.

[15] Brys, G., Hubert, M., Strayf, A. A robust measure of skewness. Journal of Computational and Graphical Statistics 2004; **13**: 996-1017.

[16] Anbarasi M. S. et al, "Outlier Detection for Multidimensional Medical Data", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. **2**(1), 2011, 512-516.

[17] Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). ROBFCA: A new approach to robust principal component analysis. Technometrics, **47**, 65-79

[18] Awodey, S. (2006). Category Theory. Oxford Logic Guides **49**. Oxford University Press.

[19] Tukey, J.W. Exploratory data analysis. Addison-Wesely, 1977.