# Matter for Debate

## Definitions for outliers in two-dimensional and higher-dimensional data

PAUL BELCHER

Using one of the usual definitions of an outlier for a set of data, this Article investigates possible definitions for outliers when the data is two-dimensional. It then considers higher-dimensional data.

Probably the most commonly used definition for an outlier is given by the definition below.

*Definition* 0. A data point $x_k \in \mathbb{R}$ is an outlier of the data set $\{x_k\}$ if $x_k$ is *either* greater than the upper quartile plus 1.5 times the interquartile range *or* less than the lower quartile minus 1.5 times the interquartile range.

There are other commonly accepted definitions of an outlier. Another popular one is that the data point is more than twice the standard deviation away from the mean.

Now suppose that we have two-dimensional (bivariate) data and want a definition to test if a data point is an outlier. We could, for example, have the two-dimensional coordinates of all the buildings in a village and wish to test if any buildings could be considered as outliers. This situation is illustrated in Figure 1.

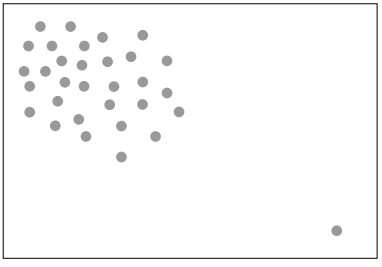

FIGURE 1

Let the two dimensional data be represented by $\{(x_i, y_i, i \in \mathbb{Z}^+, 1 \leqslant i \leqslant m, m \in \mathbb{Z}^+)\}$.

Two definitions that could be considered for the data point $(x_k, y_k)$ to be an outlier are:

*Definition* 1: The data point $(x_k, y_k) \in \mathbb{R}^2$ is an outlier for the data set $\{(x_k, y_k)\}$ if *either* $x_k$ is an outlier for the set $\{x_i\}$, by Definition 0, *or* $y_k$ is an outlier for the set $\{y_i\}$, by Definition 0.

*Definition* 2: The data point $(x_k, y_k) \in \mathbb{R}^2$ is an outlier for the data set $\{(x_k, y_k)\}$ if $x_k$ is an outlier for the set $\{x_i\}$, by Definition 0, *and* $y_k$ is an outlier for the set $\{y_i\}$, by Definition 0.

https://doi.org/10.1017/mag.2024.119 Published online by Cambridge University Press

Consider the scatter diagram shown in Figure 2 showing 23 two-dimensional data points.

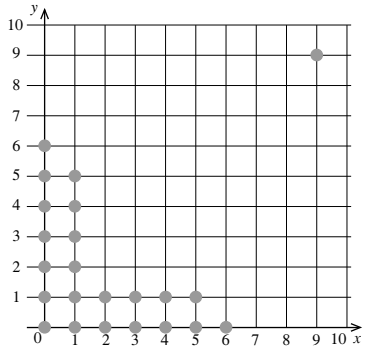


FIGURE 2

Due to the symmetry, the five-number statistical summary will be the same for the set of 23 $x$-values as it is for the 23 $y$-values and is:

$$\min = 0 \qquad Q_1 = 0 \qquad Q_2 = 1 \text{ (median)} \qquad Q_3 = 4 \qquad \max = 9$$

The interquartile range is 4 and as $4 + 1.5 \times 4 = 10$ neither the $x$-data nor the $y$-data has any outliers.

Hence the data point (9, 9) would not be an outlier according to either Definition 1 or Definition 2, but visually we naturally think that it should be considered as an outlier. Hence Definitions 1 and 2 will be considered as unsatisfactory. There are other examples of two-dimensional data sets that could be given to reinforce this decision.

To apply Definition 0 of an outlier for one-dimensional data to two-dimensional data we could transform the two-dimensional data to one-dimensional data. Let the centre of the set of data be defined by $C\,(m_x, m_y)$ where $m_x$ is the median of the $x$-data and $m_y$ is the median of the $y$-data. (It would have been possible to define $C$ by $C = (\bar{x}, \bar{y})$ where $\bar{x}$ is the mean of the $x$-data and $\bar{y}$ is the mean of the $y$-data. However the fact that the median is a more robust statistic than the mean explains why the median has been chosen.) For each data point $(x_i, y_i)$ the distance to the centre is calculated and denoted by $d_i$, which gives a set of one-dimensional data where Definition 0 can be applied. (An alternative to taking the distance between $C$ and the point $(x_i, y_i)$ would be to use the Manhattan metric and let $d_i = |x_i - m_x| + |y_i - m_y|$.) Then a new definition for the point $(x_k, y_k)$ to be an outlier could be:

*Definition* 3: The data point $(x_k, y_k) \in \mathbb{R}^2$ is an outlier for the data set $\{(x_i, y_i)\}$ if the distance $d_k$ is an outlier by Definition 0 for the set $\{d_i\}$.

This definition would certainly be suitable for the type of data demonstrated by Figure 1.

https://doi.org/10.1017/mag.2024.119 Published online by Cambridge University Press

Let us apply Definition 3 to the data given by Figure 2. Here $C = (1, 1)$ and the five-number statistical summary for the set $\{d_i\}$ is:

$$\min = 0 \qquad Q_1 = \sqrt{2} \qquad Q_2 = \sqrt{5} \qquad Q_3 = 4 \qquad \max = \sqrt{128}$$

The interquartile range is $4 - \sqrt{2}$ and as $4 + 1.5 \times (4 - \sqrt{2}) < \sqrt{128}$ this definition (pleasingly) gives that the point (9, 9) is the only outlier.

There could be data sets where this definition does not work well. A proposed rule of this type is just a guide and considering the scatter diagram visually might often be a better guide. The decision as whether or not to consider a data point as an outlier is often a matter of judgement based on experience and understanding of the context.

Let us now use Definition 3 and consider outliers in another set of two-dimensional data. We obtain an interesting result. The data is

$$\left\{(0, 0), \left(\cos\frac{j\pi}{11}, \sin\frac{j\pi}{11}\right) \text{ where } j \in \mathbb{N},\ 0 \leqslant j \leqslant 21\right\}$$

and is shown in Figure 3.

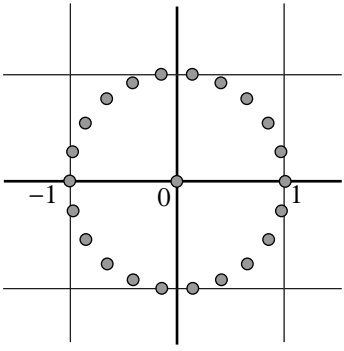


FIGURE 3

Due to the symmetry the centre $C = (0, 0)$. All of the $d_i$ will equal 1 with the exception of the one corresponding to the data point (0, 0) which will equal 0. Hence the interquartile range will equal 0 and by Definition 3, there is exactly one outlier which is the point (0, 0). It is noted that this point would not be an outlier under Definitions 1 or 2. It might seem strange that the centre can be an outlier and that the outlier is an 'insider' but visually this makes sense.

The concept of Definition 3 can be carried over to three-dimensional data. We could have a galaxy of stars and wish to determine if a particular star was to be considered as an outlier. Figure 1 could be viewed three-dimensionally to represent this. The three dimensional data would be transformed to one-dimensional data and then Definition 0 applied. The Centre would be the triple consisting of the $x$, $y$, $z$ medians and the $d_i$ would be the distances from each point to the centre.

https://doi.org/10.1017/mag.2024.119 Published online by Cambridge University Press

Similarly Definition 3 could be used for $n$-dimensional data. Letting $l$ pieces of data be represented by

$$\left\{\left(x_{1,i}, x_{2,i}, x_{3,i}, \ldots, x_{n,i}, , i \in \mathbb{Z}^+, \ 1 \leqslant i \leqslant l, l \in \mathbb{Z}^+\right)\right\},$$

the centre will be $(m_1, m_2, m_3, \ldots, m_n)$ comprising the medians, and distances will be given by $d_i = \sqrt{\sum_{j=1}^{n} (x_{l,j} - m_j)^2}$. For an example of five-dimensional data we could consider data points as $(x_i, y_i, z_i, t_i, T_i)$ representing a body at position $(x_i, y_i, z_i)$ at time $t_i$ and with temperature $T_i$.

Another way of transforming two-dimensional data to one-dimensional data would be, for each data point, to let $h_i$ represent the distance from that point to its nearest neighbour. This would lead to another definition

*Definition* 4: The data point $(x_k, y_k) \in \mathbb{R}^2$ is an outlier for the data set $\{(x_i, y_i)\}$ if the data point $h_k$ is an outlier by Definition 0 for the set $\{h_i\}$.

For data of the type given by Diagram 1, Definition 4 would be fine. For the data given by Diagram 3 it would also conclude that the point (0, 0) was an outlier. For the data given by Diagram 2 it would also conclude that the point $(9, 9)$ was an outlier. However if, with the data from diagram 2, the point (0, 0) was replaced with the point $(10, 10)$, then all the $h_i$ would be equal to 1 and Definition 4 would not give any outliers. Visually in this case it looks as though both (9, 9) and $(10, 10)$ should be outliers and Definition 3 does indeed give these two points as the outliers. Hence Definition 4 will not be considered any further.

The generalisation of Definition 3 to $n$-dimensional data should not be reduced down to one-dimensional data instead Definition 0 should be applied directly. This is illustrated in the following example. Let the one-dimensional data be:

2 9 9 10 11 13 14.

This has 2 as the only outlier. Applying Definition 3 the centre would be considered as $(10)$ and the distances would be given as

8 1 1 0 1 3 4.

This data has an interquartile range of 4.5 and so would not have any outliers. Definition 3 causes a 'folding over' effect that we do not want.

There are other possible methods for deciding how to define outliers for $n$-dimensional data. For example for each data point, find the median (again an alternative would be to use the mean) of the distances to all of the other data points and denote this by $a_i$. So this, in the two-dimensional case, would lead to a further defintion

*Definition* 5: The data point $(x_k, y_k) \in \mathbb{R}^2$ is an outlier for the data set $\{(x_i, y_i)\}$ if the data point $a_k$ is an outlier by Definition 0, for the set $\{a_j\}$.

This Definition 5 would work well for data of the type shown in Diagram 1.

https://doi.org/10.1017/mag.2024.119 Published online by Cambridge University Press

The reader is invited to propose their own definition for an *n*-dimensional data outlier.

The author's present preferred one is Definition 3.

Another point that is worth mentioning with outliers is that just because a data point has been identified as an outlier does not necessarily mean that this data point should be removed from the data set. A mistake could have been made and this is worth checking. However it could have been a very genuine piece of data that is just atypical of the rest and there would be no justification in removing it.

*Acknowledgement*

The author wishes to thank the referee for helpful comments particularly with regard to the layout of the definitions.

10.1017/mag.2024.119 © The Authors, 2024
Published by Cambridge University Press
on behalf of The Mathematical Association

PAUL BELCHER
*49 Main Road*,
*Ogmore-by-Sea, Bridgend*,
*Vale of Glamorgan CF32 0PL*
e-mail: *paul.belcher26@gmail.com*