Islamic University of Technology (IUT)

Organization of Islamic Cooperation (OIC)
Department of Electrical and Electronic Engineering (EEE)

Course No.	Math 4522
Course Name	Numerical Methods Lab.
Experiment No.	08
Experiment Name	Outlier Analysis

Objective

To get familiarized with the concept of outliers and their detection techniques.

Outliers

In statistics, an outlier is a data point that differs significantly from other observations. They are expected to occur due to natural variations. An outlier may be due to some experimental errors or scribing mistakes. But they are not necessarily bad data. For instance, in the case of credit card transactions, an outlier may represent fraudulent activity. Whereas in the case of patients age documentation, an age higher than 100 or negative value would mean some mistakes. Therefore, it entirely depends on the attributes of the data whether a value is an outlier or not. In some cases, an outlier may represent some significant characteristics, whereas in other cases, it may be just a wrong entry, which can cause serious problems in statistical analysis. In the regression analysis, you have seen the effect of outliers and how they can bias the model. Not only in regression but also in other predictive modeling tasks, or unsupervised learning tasks, outliers may have some adverse effects. It is often termed a "silent killer" due to that. Therefore, it is necessary to understand how to detect outliers and how to treat them effectively.

How to identify outliers?

The first thing you need to understand is that there is no universal way of identifying outliers. It depends on the case or data characteristics. Let’s elaborate on this issue.

Credit/Debit card fraud is a serious issue as it can cause significant loss to an individual or to the bank. Therefore, identifying such fraudulent activity is quite important. A bank generally has their own security system to detect such issues. They can temporarily stop transactions from a card to

Prepared by

ASIF NEWAZ
Lecturer, Department of EEE, IUT

avoid such scenarios. It is mainly based on the concept of pattern recognition. Let’s look at an example to explain that. All the students in IUT are given a debit card for transactions/money withdrawal. We can safely assume that, usually, the withdrawal amount of the students is less than 10000 taka. It is unlikely for the students to withdraw, say, 30000 taka in one transaction from the atm booth. Therefore, in the context of student transactions, an amount like that can be considered an outlier. However, the bank provides the same card to other users as well. Their usual transactions may or may not be higher than that. So, in the context of the general population, should 30000 taka be considered an outlier? Should the value be lower or higher for accurate fraud detection?

Methodologies for discerning outliers

There are many different approaches to identifying outliers. The efficacy of these approaches depends largely on the distribution of the data. These techniques can be broadly classified as univariate and multivariate. Some of the techniques are mentioned below.

i. Z-score => $\mu \pm 3\sigma$ (μ=mean and $\sigma = standard\ deviation$)
ii. Median => $med \pm 3c * MAD$ (MAD= Median Absolute Deviation)
iii. Inter Quartile Range (IQR) => Q3+1.5*IQR or Q1–1.5*IQR
iv. Percentiles => p>95%
v. Hypothesis testing
vi. Sliding Window technique (particularly useful for time-series data)
vii. Visualization methods – Box plot, QQ plot, Swarm plot, etc.

Let us discuss some of these techniques –

Z-score

Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. It is measured in terms of the standard deviation.

$$z_i = \frac{x_i - \mu}{\sigma}$$

A z-score of 0 means that the value is exactly equal to the mean value of the distribution. Any value that falls outside the range of 3σ is considered an outlier in this approach. This works well when the data is normally distributed. However, it is not acceptable for skewed distributions or data with large outliers.

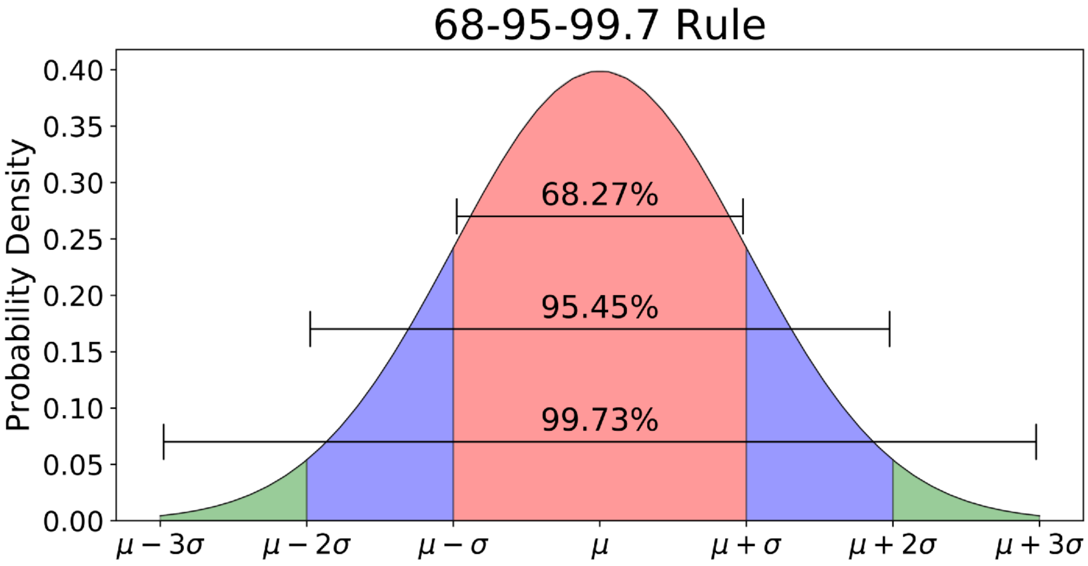
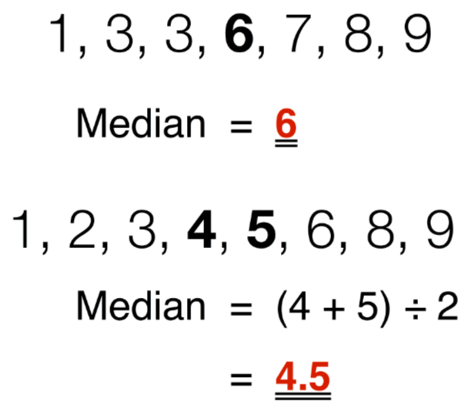


Fig. 1. Z-score

Median

The median is calculated by arranging the values in numerical order, dividing the total number of values by two, then rounding that number up (if there are odd number of values) or, by averaging the number in that position and the next position (if there are even number of values). Any values that fall outside the range of $med \pm 3c * MAD$ are considered outliers in this approach.



$$1, 3, 3, \mathbf{6}, 7, 8, 9$$

$$\text{Median} = \underline{\underline{6}}$$

$$1, 2, 3, \mathbf{4}, \mathbf{5}, 6, 8, 9$$

$$\text{Median} = (4 + 5) \div 2$$

$$= \underline{\underline{4.5}}$$

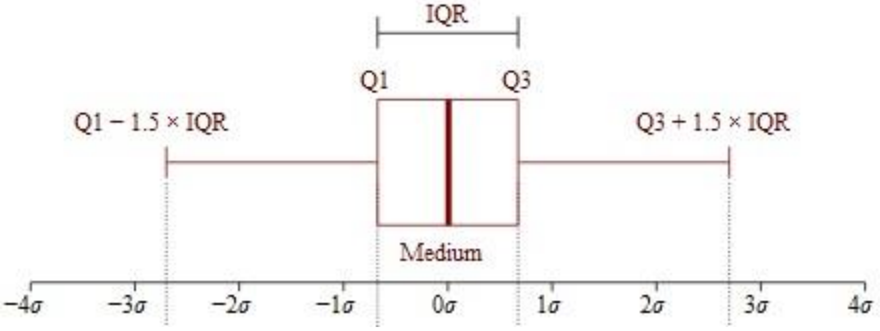
Prepared by

ASIF NEWAZ
Lecturer, Department of EEE, IUT

Take a look at the following example: say, you are calculating the average temperature of 10 objects in a room in degree Celsius and the values are as follows – 20, 25, 24, 22, 23, 28, 26, 180, 30, and 21. By looking at the values, it is evident that there is clearly an object that falls outside the typical sample type. However, if you look at the mean of the distribution, it is 39.9. Does the value represent the general trend of the data? If you calculate the median, you will get 24.5, which is more representative of the data. Therefore, median is more robust to outliers and is more suitable when the data is skewed.

IQR

It stands for Inter-Quartile Range. This method is also quite robust to outliers and suitable for general distributions. Look at the following figure.



Here, Q1 is called the lower quartile, and Q3 is called the upper quartile. The difference between these two values is called the IQR. The lines extending from Q1 and Q3 up to 1.5 times the IQR are called whiskers. Any values outside that range are considered outliers. The middle point (Q2) is the median, which is a measure of central tendency.

This can be clearly visualized from the box-plot.

What to do with outliers

Once you have detected outliers, the next question is what to do with them. There are four things that you can do with the outliers –

i. Ignore
ii. Remove
iii. Replace
iv. Isolate

The right approach depends on the data distribution, the size of the data, number of outliers, the types of outliers, the predictive model, and several other things. For instance, if the dataset size is small, removing outliers may not be appropriate. Just because one feature value is wrong does not mean all the other feature values are wrong. Replacing those values with the general feature value (mean) might be more suitable. Again, if there are only a few outliers compared to the number of samples, then those data points can be removed without hampering the distribution of the data. Some predictive modeling algorithms like the Random Forest are robust to outliers. Therefore, when using such algorithms, outliers can be ignored.

MATLAB Functions

- MATLAB provides a simple function to easily identify outliers. It returns logical true or false. There are several methods that MATLAB supports. Take a look at the documentation for more information.

  tf= isoutlier (data, find_method)

- To replace the outliers with some other values, you can use the following function. There are several techniques that you can use to fill in the values. Nearest, previous, linear interpolation, etc. are some of them. Check the documentation for more information.

  modified_data= filloutliers (data, fill_method)

  modified_data= filloutliers (data, fill_method, find_method)

- To remove outliers, use the following function.

  modified_data= rmoutliers (data, find_method)

Lab Task

For the given heart disease dataset, try to find outliers using different techniques. Use box plots to visualize the results.

Prepared by

ASIF NEWAZ
Lecturer, Department of EEE, IUT