Outlier Labeling Method for Univariate Data for Module Test and Die Sort

T. Saeger, B. Kleven, I. Otero, M. Wallace and R. Ziglar

Qorvo, 2300 NE Brookwood Pkw, Hillsboro, OR 97124
e-mail: thorsten.saeger@qorvo.com

Keywords: Outlier, PAT, Medcouple, Boxplot

Abstract

Typical data sets at Qorvo, like results from packaged part test or die level test, will contain unusually large or small values when compared with others in the data set, which may or may not pass specification limits. Such outliers will have negative effects on data analyses such as cpk-Analysis, ANOVA and regressions. More importantly, outliers, particularly those passing specification limits, may indicate defective parts or pose a reliability concern. For engineering teams in Development and Production it is standard practice to identify or label outliers in their data sets, so that those can be removed to improve data analyses and to protect customers against receiving defective parts.

The aim for this work is to provide product engineering teams at Qorvo with reliable outlier labeling methods particularly for skewed distributions. While the ultimate goal is to label and to remove outliers, we also need to balance this with unnecessary yield loss for the financial bottom line.

MOTIVATION

Why is it so important to label data points or observations which are obviously different than other points in a data set as outliers? To put it simply, something is different with this packaged part or die on a wafer. Outliers, particularly those passing specification limits, may indicate defective parts or pose a reliability concern. The automotive industry uses a stronger expression "abnormal" [1], thus asks suppliers not to ship those devices. Let's look at an example. Qorvo's legacy TriQuint organization implemented a modified Part Average Testing (PAT) at final test to screen outliers in real time [2]. In December 2013 the author of this paper observed an increased rate of PAT failures for a leakage measurement on the high band input pin of a GSM/EDGE Power Amplifier Module. Subsequent data and failure analysis showed that the PAT failure is associated with a highly localized fab defect on the base contact of an ESD protection diode of the high band HBT die [3], which clearly posed a reliability risk in the field.

The second aspect is the impact outliers have on data analyses such as cpk-Analysis, ANOVA and regressions. Let's turn the view to a simple linear regression example shown in Figure 1 to illustrate the impact of outliers.

[1] The author can only speculate on why that is, but it is clearly beyond his expertise

The left chart shows a dataset of 100 records with a linear relationship between two variables with three outliers clearly visible. The same data set is shown on the right with the outliers removed. One will note the difference in slope and intercept of the fitted lines, as the line in the left chart is tilted up towards the outliers.

TABLE I
SUMMARY OF LINEAR REGRESSION (ESTIMATE ± STANDARD ERROR)

	With outliers	w/out outliers
Intercept	4.6 ± 1.3	7.5 ± 0.3
Slope	6.5 ± 0.2	5.9 ± 0.1

Table 1 summarizes the results of the linear regression applied to this data set with the three outliers included and excluded. It shows the leverage the three outliers of this dataset of 100 records have, particularly on the intercept. To be clear, outliers are important as further analysis may point to the mechanism causing their difference, e.g. a fab process or assembly issue. However, for estimating the linear relationship between the two variables they are a hindrance increasing the error of the linear model.

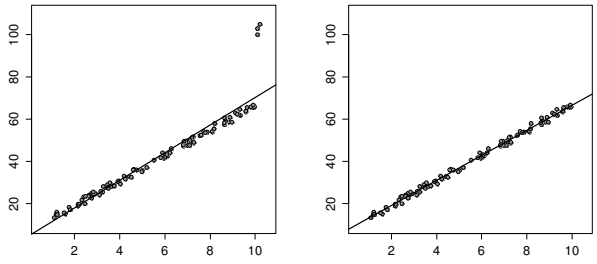

Figure 1. Comparison of a linear regression of a) dataset with outliers on the left and b) same dataset with outliers removed

WHAT IS AN OUTLIER?

The human eye is remarkable at labeling data points as outliers[1]. Going back to the left chart in Figure 1 the outliers are easy to spot. Figure 2 shows a histogram and the reader will again easily identify the main distribution (light gray) and the outliers (dark gray). Following this notion outliers are defined as follows:

Outliers are observations which are not connected to the main distribution.

While this definition is a necessary condition we also need to watch for the distribution of data points disconnected from the main distribution. If the disconnected outlying distribution is large enough relative to the main distribution we need to consider it a second mode and not outliers.

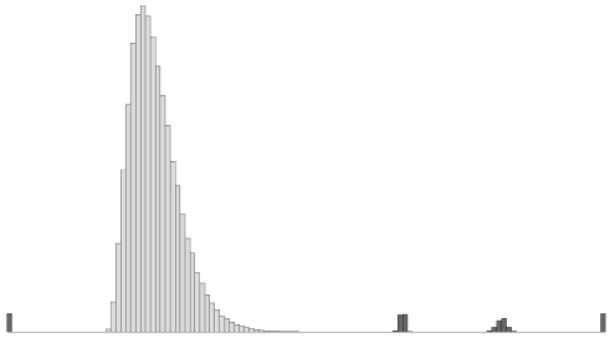

Figure 2: Histogram of a main distribution (light grey) and outlier distributions (dark grey)

Armed with that definition the quest for a reliable outlier labeling method is on. As noted, the main distribution in Figure 2 has a positive skew, which is one of the main challenges for labeling outliers. It is not uncommon that outlier labeling or identification assume a normal distribution [5]. However, by the author's experience this quite often is not the case. Other requirements for a method are availability in statistical data analysis software, execution speed and suitability for use in an automated fashion.

OUTLIER LABELING METHODS

At Qorvo's legacy TQS and RFMD Product Engineering organizations, various Outlier labeling methods have been thought of and a few had been in use. In addition, a literature search in the statistical data analysis domain had been conducted. This led to a list of 13 Outlier labeling methods.

In the following we tested all 13 methods against a typical case of a univariate distribution shown in Figure 2, the measurement of a parameter on ~ 100000 parts. As expected, the main distribution (light grey) has a positive skew. The reason for this can be multifold and won't be discussed in this paper. The outliers (dark grey) to the left and far right appear to be unimodal which may indicate clamping of the test system. The two outlying distributions next to the right of the main distribution could be indicative of a process or assembly issue. Based on the results of this test we identified 5 methods to be tested further. We selected 5 typical products ranging from PA modules to filters. The five methods have been applied to each of the measurements taken on production batches randomly selected from a 1 year time period.

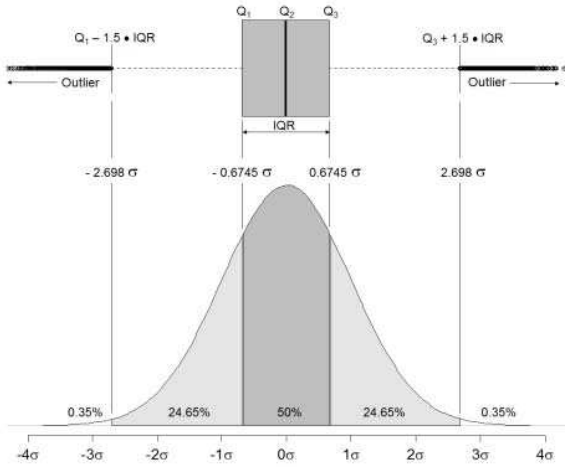

Figure 3: Relationship of a normal N(0,1) distribution to a Boxplot

This work showed clearly that any reliable outlier labeling method is based on robust statistics which we will discuss in the following.

SIGMA EQUIVALENT

For a Normal or Gaussian distribution the location and the spread can be expressed by the following robust statistics:

$$Robust\ Mean = Q_2\ (the\ median)$$

Note: For a ranked data set Q_2 is the middle data point if the sample size is an odd number. If the sample size is an even number, Q_2 is the average of the two middle points.

$$Robust\ Sigma = \frac{Q_3 - Q_1}{1.3490} = \frac{IQR}{1.3490}$$

Note: Q_1 is the point ¼ of the way through the ranked data and Q_3 is the point ¾ of the way through the ranked data.

The value of denominator of the robust Sigma for a Normal or Gaussian distribution is illustrated in Figure 3, where one can see that $Q_1 = -0.6745\sigma$ and $Q_3 = 0.6745\sigma$ leading to the definition of the robust Sigma. From that, the relationship between the Factor f of a box plot (the length of the whiskers) and the number of Sigma n_σ of a Normal or Gaussian distribution can be derived as:

$$f = \frac{n_\sigma - 0.6745}{1.3490}$$

From Figure 3 also follows how a boxplot corresponds to a Normal or Gaussian distribution. In this example the boxplot's whiskers are calculated using a factor $f = 1.5$, also referred to as the inner fence. As one can see, the inner fences equate to $\pm\ 2.698\sigma$. This would label 0.35% (or ~3488 ppm) on each side of the distribution as outliers. It has to be noted that the typical boxplot labels data points as outliers on both sides of a distribution, thus ~6976ppm or ~0.7%. From a yield perspective, this rate is too high. Furthermore, most of the observations are not disconnected from the main

distribution (see the white areas under the normal distribution in Figure 3). To conclude, the normal boxplot is a very powerful tool in exploratory data analysis [6], but should not be used indiscriminately for outlier labeling in a production environment. Table 2 gives an overview of different boxplot factors *f* and the sigma equivalent n_σ of a normal distribution and the corresponding outlier labeling rate in ppm (one-sided).

TABLE II
SIGMA EQUIVALENT n_σ, BOXPLOT FACTOR *f* AND PPM OF OBSERVATIONS OUTSIDE OF A GIVEN n_σ OR *f* FOR A NORMAL DISTRIBUTION (ONE-SIDED)

Description	n_σ	*f*	Outlier, ppm
Inner Fence	2.698	1.500	3487.872
Outer Fence	4.722	3.000	1.168
3.0 Sigma	3.000	1.724	1349.898
4.5 Sigma	4.500	2.836	3.398
6.0 Sigma	6.000	3.948	0.001

MODIFIED PAT OR BOXPLOT METHOD

In our testing we found that the modified PAT or modified Boxplot method introduced in [2] is the method of choice in automated environments like real time screening at packaged part test, which is defined as follows:

$$LOL = Q_1 - \frac{n_\sigma - 0.6745}{1.3490} \cdot IQR,\ UOL = Q_3 + \frac{n_\sigma - 0.6745}{1.3490} \cdot IQR$$

With:

LOL … Lower Outlier Limit, observations smaller are labeled as outliers

UOL … Upper Outlier Limit, observations greater are labeled as outliers

To overcome the issue with skewed or highly tailed distributions using this method, n_σ can be set asymmetrically (e.g. LOL is set to $n_\sigma = 6$ and UOL is set to $n_\sigma = 9$).

ADJUSTED BOXPLOT METHOD

In a literature search we found the work of Ellen Vanderviere and Mia Huber: "An adjusted Boxplot for skewed distributions" [8]. To overcome the issue with skewed or tailed distributions, the boxplot method can be enhanced by introducing the Medcouple *MC* as follows:

For a positive (right-side tail) skew:

$$LOL = Q1 - 1.5 \cdot e^{-4 \cdot MC} \cdot IQR, UOL = Q3 + 1.5 \cdot e^{3 \cdot MC} \cdot IQR,\ MC \geq 0$$

For a negative (left-side tail) skew:

$$LOL = Q1 - 1.5 \cdot e^{-3 \cdot MC} \cdot IQR, UOL = Q3 + 1.5 \cdot e^{4 \cdot MC} \cdot IQR,\ MC < 0$$

Replacing the boxplot factor $f = 1.5$ with the Sigma equivalent of *f* we get:

For a positive (right-side tail) skew:

$$LOL = Q1 - \frac{n_\sigma - 0.6745}{1.3490} \cdot e^{-4 \cdot MC} \cdot IQR,$$

$$UOL = Q3 + \frac{n_\sigma - 0.6745}{1.3490} \cdot e^{3 \cdot MC} \cdot IQR,\ MC \geq 0$$

For a negative (left-side tail) skew:

$$LOL = Q1 - \frac{n_\sigma - 0.6745}{1.3490} \cdot e^{-3 \cdot MC} \cdot IQR,$$

$$UOL = Q3 + \frac{n_\sigma - 0.6745}{1.3490} \cdot e^{4 \cdot MC} \cdot IQR,\ MC < 0$$

Now, we need to discuss a few things. The Medcouple MC is a robust statistic that measures skewness and is defined as follows [7]:

$$MC = med\ h(x_i, x_j),$$

$$with\ h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i}$$

$$x_i \leq Q_2 \leq x_j\ and\ all\ pairs\ x_i \neq x_j$$

A data set is now split into values x_i smaller or equal to the Median Q_2 and values x_j greater or equal to the median Q_2. The function $h(x_i,x_j)$ now measures the normalized difference between the distances of x_i and x_j to the median for all pairs $x_i \neq x_j$. With *n* values in our data set we get $(½n)^2$ $h(x_i,x_j)$ values and *MC* is the median of those. We depart here from a simple ranking of the data points to take the quantiles, which makes the modified PAT fast and suitable for real time screening as the calculation of $(½n)^2$ values will take its time. The implementation used in our investigation uses a fast algorithm shown in [7] and the impact on off-line analysis such as re-mapping wafers or strips or other data analysis tasks is not noticeable.

The Medcouple *MC* is multiplied with a scaling factor forming the exponent of the *e* functions in above equations. The derivation of the scaling factors are shown in [8] and we used those as defaults. In the case of a symmetrical distribution (no skew or tails), the *MC* will be 0 and above equations will simplify to our modified PAT.

A COMPARISON

In this work we could confirm that the modified PAT algorithm [2], is the method of choice for dynamic screening at package part test. Any tailed or skewed distribution can be addressed by setting an asymmetric sigma equivalent for the lower and upper outlier limit calculation. For any other scenario, like data analyses and outlier removal on wafers following die sort or strip test, we found a reliable method in the adjusted boxplot method, modified in the same manner as the modified PAT. But how do those methods compare when applied to the data set from Figure 2?

Assume we need to label outliers with a Sigma Equivalent of n_σ=3. Applied to both sides of the distribution we would expect that all disconnected observations and approximately 2700 ppm observations from the main distribution would be labeled as outliers.

In this exercise we added the classic Mean ± n_σ *Sigma method as a third item. The results are summarized in Table III and visualized in Figure 4.

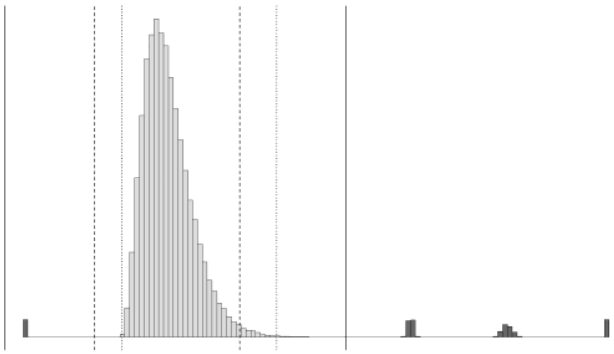

Figure 4. Outlier labeling methods applied to test distribution with $n_\sigma = 3$ (see text for details)

Table III
Comparison of three outlier labeling methods

Method	**Disconnected observations labeled as outlier**	**Outlier (two-sided), ppm**
Mean ± n_σ*Sigma	No	0
Modified PAT symmetric n_σ	Yes	~10600
Adjusted Boxplot	Yes	~1060

What the results show is that the classic Mean ± n_σ*Sigma method is not applicable (solid lines in Figure 4). The limits to label observations as outliers are too wide. It simply shows the fact that mean and standard deviation are impacted by tailed distributions and the outliers themselves. The modified PAT method (dashed lined in Figure 4) labels the disconnected observations as outliers, but as one can see the upper limit cuts too far into the main distribution labeling ~10600ppm observations as outliers which is too much compared at the expected range of approximately 2700ppm, causing unwarranted yield loss. The Adjusted boxplot methods (dotted lines in Figure 4) encapsulate the main positive skewed distribution closely as required and labeling ~1060 observations of the main distribution as outliers.

LIMITATIONS

In our investigation the adjusted boxplot method performed with the required outcome for univariate data sets. But, as one can see in the comparison summarized in Table III, it labeled only ~1060 ppm observations as outliers compared to the expected range of approximately 2700ppm. We need to be aware that any outlier labeling methods is prone to effects referred to as masking and swamping [5].

In our daily lives masking is the one effect impacting us most. Masking is the effect when true outliers mask other true outliers such that they appear as non-outliers.

Looking at the three outlying populations to the right of the main distribution in Figure 4, one can argue that the outliers to the far right are masking the two others. If those outliers are removed from the calculation the adjusted boxplot methods labels ~2100ppm of the main distribution as outliers which is within the expected range.

The last limiting effects to be considered are bimodal or multi-modal distributions. In this case the application of any outlier labeling method should only be done when results can be verified using visualizations like histograms. If feasible, the cause of the multimodality needs to be determined (e.g. wafers, epi growth runs, assembly batches, test sessions etc.). The outlier labeling method should then be applied to the sub groups individually.

CONCLUSIONS

In this work we confirmed that the modified PAT algorithm [2], is the method of choice for dynamic screening at package part test. Any tailed or skewed distribution can be addressed by setting an asymmetric sigma equivalent for the lower and upper outlier limit calculation. The calculation of outlier limits is simple and fast making it suitable for real time screening at packaged part test.

For any other scenario, like data analyses and outlier removal on wafers following die sort or packaged part strip test, we found a reliable method in the adjusted boxplot method, modified in the same manner as the modified PAT. Using a fast algorithm, available as C++ code, Matlab and R, the execution time for the calculations is not noticeable on a common laptop or desktop PC even for 1000000 observations.

ACKNOWLEDGEMENTS

The authors would like to thank George Weaver, Industrial Statistician at the Qorvo Oregon Process Engineering group for the discussion and reading of the manuscript. Additionally, we like to thank our senior director James Eastham for the support of this work.

REFERENCES

[1] Automotive Electronics Council, AEC-Q001, Guidelines for part average testing, Rev-D, December 2011

[2] Douglas Pihlaja, *Real Time Dynamic Application of Part Average Testing (PAT) at Final Test*, 2013 CS ManTech Technical Digest, pp. 59-61, May 2013.

[3] Thorsten Saeger, TriQuint Oregon - *Internal Report*, December 2013

[4] Neill Patterson, A Robust, *Non-Parametric Method to Identify Outliers and Improve Final Yield and Quality*, 2012 CS ManTech Technical Digest, pp. 165-167, April 2012.

[5] NIST Engineering Statistics Handbook, *Detection of Outliers*, http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm, retrieved on 04/15/2015

[6] J. W. Tukey, *Exploratory Data Analysis*, Addison–Wesley, 1977

[7] G. Brys, M. Huberty, A. Struyfz, *A robust measure of skewness*, Journal of Computational and Graphical Statistics, 2004

[8] E. Vanderviere, Mia Huber, *An Adjusted Boxplot for skewed Distributions*, COMPSTAT Symposium, 2004

ACRONYMS

PAT: Part Average Testing
Q_n: Quantiles with n= 1, 2, 3
LOL: Lower outlier limit
UOL: Upper outlier limit
MC: Medcouple