

# A Robust, Non-Parametric Method to Identify Outliers and Improve Final Yield and Quality

Neill Patterson

TriQuint Semiconductor, Hillsboro Oregon, USA (neill.patterson@tqs.com)

## Abstract

**This paper discusses the shortcomings of two widely-used outlier detection methods, and then introduces a more robust *non-parametric* algorithm that can better handle the wide variety of non-standard distributions seen in the semiconductor manufacturing industry. When applied to test data gathered at a wafer sort step, this algorithm allows for the removal of outlier die (that otherwise passed test limits) in order to improve quality and yield in the final product.**

**The algorithm was designed to be as simple as possible, with a minimum number of assumptions about the underlying dataset. Its design was inspired by the human ability to identify outliers almost instantaneously from almost any distribution. It will be presented first in mathematically-informal, yet easily-understood terms, with a more rigorous mathematical definition to follow.**

## INTRODUCTION

The problem of outliers has long plagued statisticians, scientists, and engineers of all disciplines. It is of no help that there is no formal mathematical definition of an outlier, and the entire endeavor of outlier identification will always remain somewhat subjective.

Perhaps the most popular outlier identification method in use today is the Six Sigma™ method developed by Motorola, with talk of process capability (Cpk) being ubiquitous throughout many industries. The Six Sigma™ method is an example of a *parametric* method that relies on assumptions about the underlying probability distribution – in this case, that it is Gaussian. Parametric methods are very powerful when the underlying assumptions are true, but are often woefully inadequate otherwise.

Methods relying on assumptions of normality are poor candidates at TriQuint Semiconductor, where highly

non-normal distributions are commonplace. And as more and more tests are added to a die or test structure, the odds of finding at least one distribution that is multi-modal, highly-skewed or heavy-tailed increase and the ability to exclude die based on methods that assume normality is highly diminished.

Another approach tried at TriQuint Semiconductor was based on Tukey's box-plot method, which defines outliers as being outside the interval  $[Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$ , where "Q" stands for "quartile" and "IQR" stands for "interquartile range". This method was modified at TriQuint to define outliers as being outside the interval  $[Q1 - 3.95 \cdot IQR, Q3 + 3.95 \cdot IQR]$ . The change to the IQR-multiplier was made so that if the underlying distribution is Gaussian, then this method yields the exact same results as using  $\pm 6\sigma$  limits, thus keeping with the spirit of the Six Sigma method™. This is an example of a *non-parametric* method that does not make any assumptions about the underlying probability distribution (for this reason, non-parametric methods are called "distribution-free"). Its main advantage over the Six Sigma method™ is that it is much more adept at dealing with non-symmetric distributions with heavy tails. It is a rather robust method with low computational complexity and suitable for use in many industries, but unfortunately it too is unable to handle the sheer variety of non-standard distributions seen at TriQuint.

## A MORE HUMAN METHOD

When shown a histogram of almost any dataset, even an average human with little-to-no training in statistics is able to identify outliers almost immediately. It is posited that the human algorithm for identifying outliers is to first identify the main distribution, presumably based on the region with the most area under the curve and then following that region out in both directions until the bin count drops to zero. In very simple human terms then, the definition of an outlier is *a data point not connected to the main probability distribution*.

A computer implementation of this algorithm is quite easy to produce. Binning the data and identifying bin counts of zero is trivial. Estimating area as a Riemann sum of bin counts between regions bounded by bin counts of zero and finding the main distribution is not exceedingly difficult either, but a very simple shortcut is achieved by making the following assumption: *that the median value is contained within the main probability distribution*. This is the one and only assumption made by the algorithm, and can be dispensed with at the cost of some computational complexity. This assumption is remarkably robust, and it takes a very extreme type of multi-modal distribution with clear separation between the modes to invalidate it. To date, such distributions have not been observed but seem likely to occur in multi-site measurements at the extreme limits of tester capability. For this reason, special rules are implemented to ensure that the sensitivity of the algorithm does not exceed the capability of the measurement.

In more formal mathematical terms, definitions for the main distribution and outliers for are as follows. An ordered set of elements  $\{x_1, x_2, \dots, x_N\}$  of size  $N$  form a group if, for a chosen  $\epsilon > 0$ ,  $x_i - x_{i-1} < \epsilon$  for all  $i \leq N$ . The main distribution is the group with the highest number of cardinal elements and outliers are simply defined as any elements that are outside of this group. By definition, outliers are at a distance of at least  $\epsilon$  from the end points of the main distribution. Note that  $\epsilon$  is the bin width, and choice of  $\epsilon$  controls the sensitivity of the algorithm.

## EVALUATION

Because the algorithm is not a function of any measure of distribution spread (e.g. variance, IQR, etc), it handles asymmetrical distributions very well compared to other methods. Figure 1 shows the left and right sides of a long-tailed distribution receiving radically different treatments, as is appropriate for this kind of distribution.

In addition to detecting random defects, the proposed algorithm has been useful in detecting recurring patterns of failing die that were not detectable using test limits alone. Such recurring regions of failing die are often caused by a particular tool, and this kind of information can be used to drive process improvements. Figure 2 shows an example of a non-random pattern of failing die that was visible across several wafers of the same type.

Figure 1 – A Typical Long-Tailed Asymmetrical Distribution

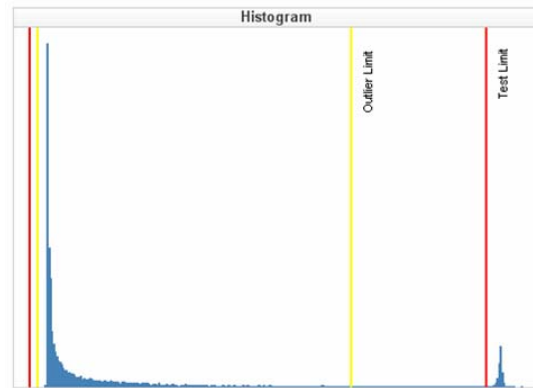
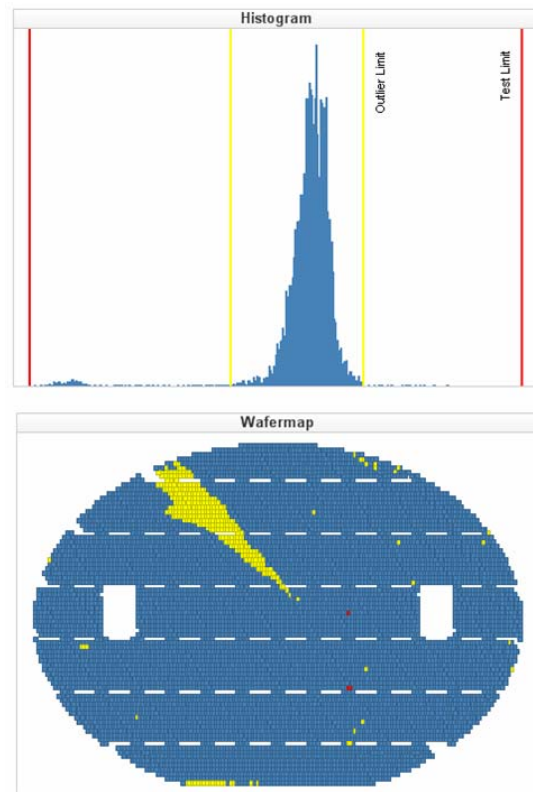


Figure 2 - A Systemic Pattern Revealed By Outlier Limits



## CONCLUSION

With a relatively small amount of computational effort, it is possible to implement a non-parametric (distribution-free) algorithm to identify outliers in a data set. This algorithm can be used solely for data analysis purposes, or can also be used to actively eliminate outlier die at wafer sort and a corresponding improvement in final yield and quality can be expected.