Statistics in Semiconductor Test: Going beyond Yield

W. Robert Daasch and C. Glenn Shirley
Portland State University

Amit Nahar
Texas Instruments

Editor's note:
The quantity and complexity of data generated at each test manufacturing step can be daunting. This article, which emerged from a tutorial presented at ITC 2008, explains the application of statistics to help process that data and provides examples of how test has shifted from descriptive to predictive methods.
—Nur A. Touba, University of Texas

As test has evolved, it has added to its original back-end role of separating good chips from bad chips by comparing measured values to preset limits. Today, test classifies chips for future use, controls and adaptively modifies manufacturing processes, and is a key input in next-generation process development and optimization. The new roles and applications require generation, transmission, storage, and processing of large amounts of data and its use in automated or manual decision-making. When the data is acquired for an individual unit, there is little additional overhead to retain data for entire wafers, lots, and so on, and this is the first opportunity to apply statistics. Once testers are connected to a data automation infrastructure and data validation controls are in place, large volumes of data can be acquired and integrated across different test flows and multiple products, from multiple factories, and over long time spans. Leveraging advanced test data integration makes it possible to apply the most sophisticated data mining and optimization methods.

For test engineering, this new frontier is called *statistical test*. The new data-intensive test applications place a premium on statistical methods capable of reducing data to a useful form for decision-making and optimization.[1] Fortunately, test engineering does not need to invent new statistical methods from scratch every time. Many statistical methods not widely known or used within test engineering are well-established for other disciplines such as agricultural sciences, genetic sciences, and manufacturing. To continue the growth of statistical test engineering, the challenge is to find and understand appropriate statistical methods, modify them correctly, and apply them to problems in semiconductor testing.

This article demonstrates the use of statistics and the application of statistical methods to problems arising in test. We do not cover the full range of statistical test but provide concrete examples. The selected applications—inference, modeling, and data mining—all emphasize how statistics are key inputs to test-engineering decision-making. The examples demonstrate combinations of familiar models such as 1-of-*k* multinomial experiments and data statistics such as χ^2 (chi square) tests of contingency tables.

Characteristics of statistical test

Statistics is the science of collecting, organizing, and interpreting experimental data to draw conclusions about unknown populations observed through experiment. Figure 1 summarizes the relationship between statistical decision-making, experiments, and populations.

In the broadest terms, statistical decisions either describe statistical properties of observation data or make statistical predictions about future events. Experimental data either helps determine parameters in statistical models or is used directly in statistical decision-making. Traditionally, test engineering has used descriptive statistics, computed from data, in statistical process control of semiconductor manufacturing. In the past 10 years, the use of semiconductor test statistics has increasingly shifted to the prediction (inference) of future events. Once largely the domain of component reliability, statistical prediction has become an integral part of volume screening of semiconductor chips.

0740-7475/09/$26.00 © 2009 IEEE Copublished by the IEEE CS and the IEEE CASS

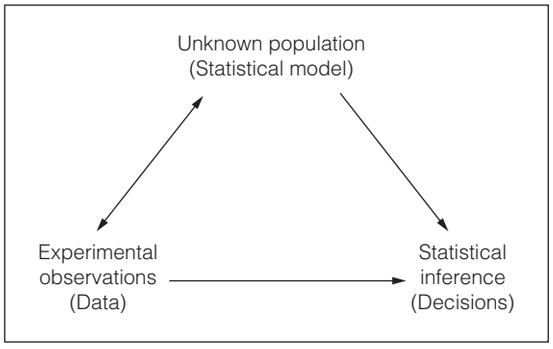

Figure 1. Statistical methods enable inferences (decisions) to be made from statistical models or directly from data. Statistical models are derived from data or sometimes simply generated from hypotheses about random or deterministic behavior.

Probabilities, distributions, and other basic statistical ideas are mathematical objects that can be manipulated according to mathematical rules. To draw conclusions and make decisions from the IC test response data, practitioners must shift their focus from the mathematics of distributions and probabilities to the statistics of the IC test response. Specifically, in test engineering, the role of statistics is decision-making. Answers to questions such as "What limits are best suited to reach the goal of reducing the ratio of test overkill to test escapes?" are experimental inferences, and statistics provide essential insights to compute and judge such answers.

A discussion about statistical tools requires basic definitions and a few key statistical ideas. The statistical literature is rich with ideas, techniques, and deep mathematics. Readers with an interest or need can find valuable resources in textbooks, monographs and topical series, and online collections.[2,3] For example, JSTOR (http://www.jstor.org) has created a high-quality, interdisciplinary archive of scholarship and is actively preserving more than 1,000 academic journals in both digital and print formats.

The basic definitions of mean and median, variance, and quartile ranges are familiar to engineers. Likewise, concepts such as the distribution, the shape of the graph, and the difference between a continuous and discrete variable are common in test engineering. These ideas are central to descriptive statistics. Less frequently used are a function's effects on a random variable, rank and order statistics, and the difference between correlation and statistical independence. The least-familiar elements of statistics are the ideas used to predict outcomes: so-called inferential statistics. Predictions are sensitive to the experimental design, which can minimize the impact of nuisance (unseen) effects. The central idea that is often misunderstood, or entirely missed, is the statistical hypothesis and the difference between the null hypothesis and the alternative hypothesis.

In this article, two applications illustrate a few statistical methods that are especially useful in statistical test. These statistical methods are as follows.

First, *define purely random, and purely deterministic, extreme scenarios*. Often, with no other information, models of statistical performance indicators can be written down for two limit cases: random and deterministic scenarios. If the predictions of these extreme cases meet the requirements, then more elaborate performance modeling is unnecessary. For example, many test outcomes can be described as a 1-of-*k* random scenario and modeled with the multinomial distribution. This article demonstrates this idea by using multisite testing and a binomial distribution—a special case of a multinomial distribution.

Second, *employ Monte Carlo (MC) synthesis*. Statistics can be a dauntingly mathematical subject, but with a laptop computer it's possible to perform useful MC simulations. One key requirement, which might be unfamiliar to engineers, is generating MC variables with prescribed degrees of correlation. The multisite-testing example demonstrates a method for synthesizing wafers in which failed dies exhibit varying degrees of spatial correlation across the wafer.

Third, *build statistical models from functions of random variables*. Test is an inherently hierarchical process in which a probability density function (PDF) modeling one element (for instance, a single test) is easily written down. In "what if" scenario modeling, a PDF is needed for a combination of many such elements. For example, to model wafer test times, a test program test time PDF is needed, and it might include the maximum, the minimum, or the sum of the assembled single tests. This article shows how the distribution of maximum test time (an order statistic) is particularly useful in modeling multisite test times.

Fourth, *define a null hypothesis and test for deviations from it.* Data mining examines data records for subtle, but potentially valuable, patterns and information. The χ^2 test for independence is a common

method for detecting correlation in attribute data organized into contingency tables.[2] In the data mining example, contingency tables reduce large volumes of wafer sort test data into a compact single metatest to predict post burn-in test response.

Modeling multisite test time

Parallel testing of dies at the wafer level is clearly an opportunity for test time reduction.[4] But, to quantify the potential benefit, a statistical model that characterizes the die-to-die variation of test time (TT) is needed. Generally, the test time of defect-free (good) dies (TTG) does not vary much, and it is usually (but not always) the longest die TT encountered. The test time of defective (bad) dies (TTB) is variable, since there are many ways for a die to be defective. Usually, TTB is shorter than TTG unless, for example, the test program has significant diagnostic tests in the failure flows of the program. For a group of dies tested in parallel (a multisite), the TT of the multisite is the longest TT of the individual dies in the multisite. If the multisite has at least one good die, this longest die TT is usually the (die) TTG. Furthermore, the frequency of occurrence of multisites with multisite TT equal to die TTG depends on the spatial distribution of good dies across the wafer. Statistical theories of combinations of functions of random variables, such as *order statistics*, can be used to derive scaling laws, which are valuable in test modeling. In the multisite example, the theoretical results from order statistics provide a scaling model for die-average TT in multisite testing.

In terms of the method shown in Figure 1, we can infer model parameters to sufficient precision from the data. That is, individual die TT distributions and the spatial distribution of defective dies are sufficiently well established. The focus in this example is on building a statistical scaling model to quantify the effect of multisite size, yield, and spatial distribution on die-average TT.

For the spatial-distribution model, a key observation is that "sufficiently well established" can in some cases mean "does not need to be known at all." In effect, for some choices of model parameters, the multisite results can be bracketed between two extreme cases, for which the results scarcely differ: *random* and *clustered* (deterministic) spatial distributions. This is an important technique in statistical modeling of test because the extreme cases are often easy to compute analytically, provided we are familiar with the appropriate statistics. The concepts of binomial statistics and order statistics provide analytical statistical models for extreme cases of randomly distributed dies and fully clustered dies on a wafer. Other examples include working out predictions of perfectly correlated tests versus completely uncorrelated tests, which can further bracket the relevant figure of merit.

When the random and deterministic cases do not sufficiently bracket the relevant figure of merit, it is still possible to avoid complicated statistical mathematics if we exploit MC simulation. In this case, it is often necessary to comprehend degrees of correlation. The multisite example shows MC simulation of degrees of spatial correlation.

Implementing multisite testing requires using a test program with at least one part that can be parallelized. For simplicity, the example assumes that the entire test program is parallelizable. For example, a BIST test could be initiated on each die, and then run without tester intervention except for monitoring a pass/fail bit. The parameters of the model are

- *multisite count* N (the number of dies tested in a single prober touchdown);
- *single-die TT distributions* for the TTG and the TTB;
- *yield* (just as for serial tests, if the TTG is the longest TT, then the more good dies are tested, the longer the average TT); and
- *degree of clustering of dies* (if all the bad dies are in one region of the wafer, then a relatively large proportion of multisite probe touchdowns will have all bad dies in them and have test times shorter than the TTG; for randomly distributed dies, more multisite touchdowns will have good dies in them, so the average test time will be longer).

TT distribution model

Assume a test program that, when run on a single die, has the TT model shown in Figure 2. In this model, the TTG is a constant, g_1, and the TTB has an exponential distribution with mean (and standard deviation) b_1. Notice that in this model the TTB can sometimes exceed the TTG. This might occur, for example, if a failure prompts a change in test program flow to include significant diagnostic tests.

During test, each multisite samples i values from the TTG distribution and $N - i$ values from the TTB distribution. The TTB for the multisite is the largest TTB of the $N - i$ values of the individual die TTBs sampled. This largest TTB is the n-of-n order statistic, with $n = N - i$, and itself defines a distribution. For the exponential distribution used as the TTB model, the average of the distribution of the n-of-n statistic is

$$b_n = b_1 H_n \tag{1}$$

where H_n is the partial sum of the harmonic series

$$H_n = \sum_{i=1,\, n} \frac{1}{i} \tag{2}$$

Figure 2 shows order statistics of the TTB for several values of n. We also need the order statistic of good dies. Since the TTG is a delta function, all samples of good dies of size n have the same TT, so the mean of the n-of-n order statistic of the TTG is

$$g_n = g_1 \tag{3}$$

Multisite TTs

All average TTs for a multisite lie between the case for which bad dies are randomly distributed across the wafer (random) and the case for which all the bad dies are in a single contiguous group (fully clustered).

Random. If the proportion of all good dies is Y, and they are distributed randomly across the wafer, then the probability p of each multisite configuration of good and bad dies is given by a term in the multinomial (in this case, binomial) expansion of $[Y + (1 - Y)]^N$. For the ith term, this is

$$p(N, i, Y) = \frac{N!}{(N-i)!i!}(1-Y)^{N-i}Y^i \quad (i = 0, N) \tag{4}$$

The TTs for each term are related to the order statistics just described. Combining them yields a model for the TT for the fully parallel test of a N-site multisite as

$$t_{\text{random}}(N, Y, b_1, g_1) = (1-Y)^N b_N + \sum_{i=1,\, N-1} p(N, i, Y) \max(b_{N-i}, g_i) + Y^N g_N \tag{5}$$

where b_n is the n-of-n order statistic of b and similarly for g. Using the model of Figure 2 and defining a TT

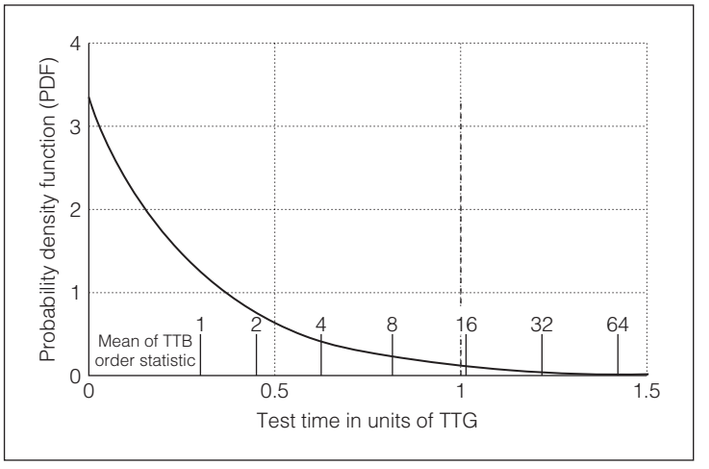

Figure 2. Test time probability density function (PDF) model. For individually tested dies, every good die has a test time good (TTG) of g_1 (dashed line), and bad dies have a test time bad (TTB) distributed as an exponential distribution (solid curve) with mean b_1. The means of various order statistics (b_1, b_2, b_4, ...) of the TTB distribution are shown. The graph is drawn for $g_1 = 1$ and $b_1 = 0.3$ (that is, $r = 0.3$, where r is b_1/g_1).

in units of g_1, Equation 5 becomes

$$t_{\text{random}}(N, Y, r) = (1-Y)^N r H_N + \sum_{i=1,\, N-1} p(N, i, Y) \max(rH_{N-i}, 1) + Y^N \tag{6}$$

where $r = b_1/g_1$. Two special cases are of interest: First, for single-die testing ($N = 1$), Equation 6 reduces, as expected, to a simple yield-weighted average of the average of the TTB and the TTG:

$$t_{\text{random}}(N = 1, Y, r) = (1-Y)r + Y \tag{7}$$

If the average of the longest TTB in a multisite is always less than the TTG, then for the model of Figure 2,

$$rH_n < 1 \quad \textit{for} \quad 1 \le n \le N-1 \tag{8}$$

Therefore,

$$t_{\text{random}}(N, Y, r) = (1-Y)^N r H_N + [1 - (1-Y)^N] \tag{9}$$

Fully clustered. In this case, we ignore multisites on the boundary between the area of bad and good dies to simplify the model. Without these cases, the TT is either the TTG (all parts pass) or the n-of-n order statistic of the TTB distribution, each in the proportion to the site's yield. Using the same scaling, the fully clustered TT becomes

$$t_{\text{clustered}}(N, Y, r) = (1-Y) r H_N + Y \tag{10}$$

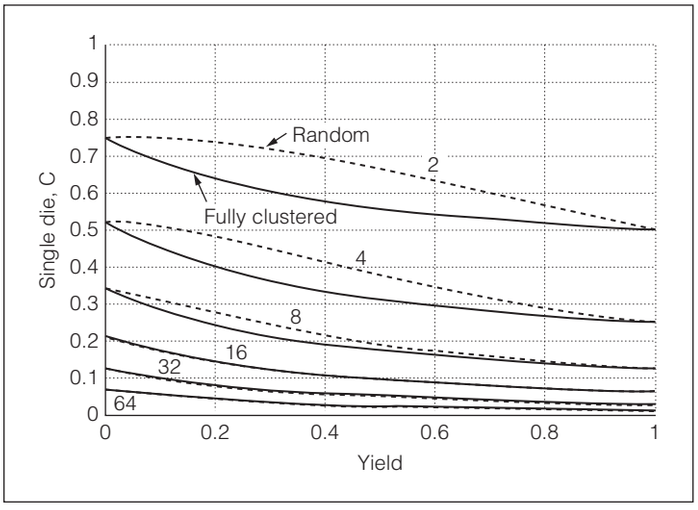

Figure 3. Average test time per die for a multisite as a function of a fraction *C* of good dies (yield) divided by average test time for a single die, for multisite sizes 2, 4, 8, ... The figure also shows clustered (solid curves) versus random (dashed curves) distributions of failing dies on the wafer. Figure 2 gives the test time model.

TT benefit of multisite test

To make an informed decision, test engineers need a figure of merit. A useful figure of merit is the average TT per die of the multisite divided by the TT for a die tested singly:

$$C = \frac{t(N, Y, r)}{N \times t(1, Y, r)} \tag{11}$$

where TTs are given by Equation 9 or 10. Smaller values of C are better. Figure 3 shows C plotted as a function of Y for several multisite sizes for the TT distributions given in Figure 2 ($r = 0.3$). When $Y = 1$, all TTs are the TTG with a single value, g_1, so the multisite TT per-die benefit is $1/N$, as expected. However, for smaller values of Y, the multisite TT benefit is reduced (C is larger) due to the effect of the TTB order statistic. This effect is greater for the random case than the clustered case. As the multisite count increases, the difference between the clustered case and the random case vanishes.

This example shows how the order statistic provides a scaling model for the multisite TT benefit, and how consideration of details of spatial correlation might be unnecessary for making decisions about implementing multisite testing. If the benefit for the random case is enough, implementation is a "go."

Intermediate degrees of clustering

As the TTB becomes smaller relative to the TTG, the area between the curves for clustering and random cases in Figure 3 expands. Hence, it might be useful to develop models for average TT per die as a function of intermediate degrees of clustering. In such cases, analytical expressions like Equation 4 for the probability of occurrence of each multisite configuration $p(N, i, Y)$ cannot easily be written down. However, it is possible to synthesize pass/fail patterns with various degrees of clustering. Figure 4 shows synthesized patterns for random and clustered patterns with $Y = 0.5$, as well as a pattern with an intermediate degree of clustering with $Y = 0.8$. We generated the patterns in Figure 4 using an algorithm that populates a lattice with failed dies according to a stochastic process, in which the probability of a site failing is influenced by previously failed neighbors. If a previously failed site has no influence on a passing site's probability of failure, the pattern generated is random. On the other hand, if a failed site makes neighboring passing sites more likely to fail, a clustered pattern is generated.[5] The range of the clustering influence modulates the degree of diffusion of the clusters. The process continues until a desired yield (fraction passing) is reached.

MC simulation is a valuable tool in statistical test engineering because it allows simulation of the manufacturing environment before silicon is available. For example, we can combine the per-die TT distribution of Figure 2 with the synthesized wafers of Figure 4 to provide a complete statistical model of multisite TT. The multisite pattern defines a stepping template across the synthesized wafer. For example, Figures 4b, 4d, and 4f show the yield distributions of the 16 sites of a 4×4 multisite for the clustering patterns in Figures 4a, 4c, and 4e, respectively. In the model just discussed, the multisite probabilities for the synthesized data (solid bars) in Figure 4 replace the binomial probabilities $p(N, i, Y)$ in Equation 4, which are shown for comparison in Figure 4 as hatched bars. Once silicon is available, the multisite template can be scanned across real, not synthesized, wafer maps to extract experimental binomial probabilities, and hence estimate the experimental values of the figure of merit.

The most important, and nontrivial, "trick" in MC simulation is to generate patterns with controlled degrees of correlation that can be varied. This applies not only to spatial patterns but also to other attributes.

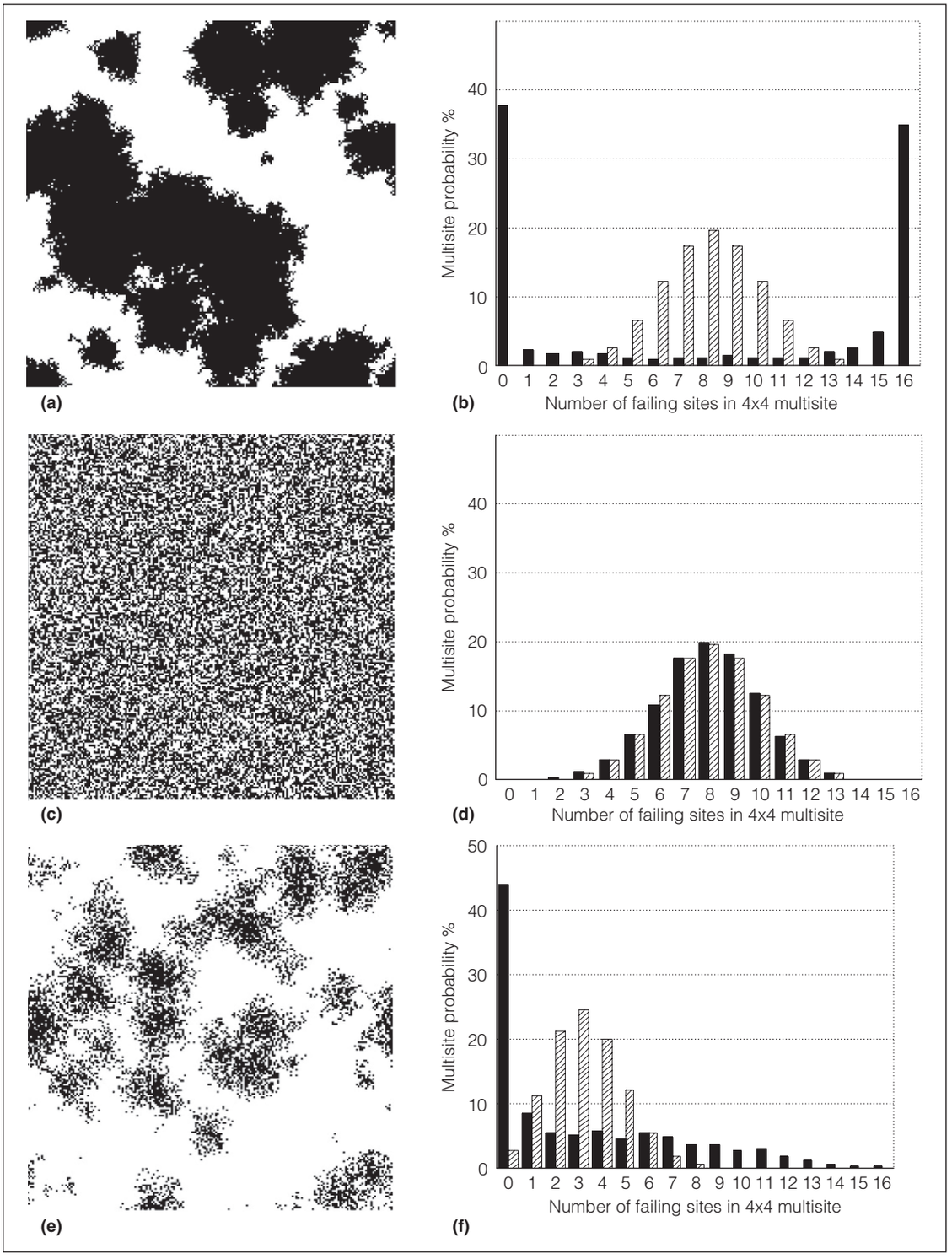

Figure 4. Synthesized probability patterns of a 4 × 4 multisite with various degrees of clustering: a pattern strongly clustered with a yield of 50% (a) and its corresponding yield distribution (b); a random pattern with a yield of 50% (c) and its corresponding yield distribution (d); and a pattern having an intermediate degree of clustering with a yield of 80% (e) and its corresponding yield distribution (f). Solid bars indicate multisite probabilities for the synthesized data; hatched bars indicate the binomial probabilities *p*(*N*, *i*, *Y*) from Equation 4.

Statistical test data mining

The central ideas of test outliers are that healthy dies have a certain intrinsic response distribution, and bad dies depart from the intrinsic healthy die distribution in detectable ways.[6] Outliers can be identified effectively when the overlap is minimized between the good- and bad-die distributions. In practice, parametric test responses such as raw I_{DDQ} show considerable overlap between healthy and faulty distributions, making it difficult to identify outliers. As technology continues to scale, this overlap is increasing.[7] One approach to reduce the overlap is to build a data-driven response model that decreases the two new distribution variances while preserving the difference between the means of those distributions.[6] In many cases, data-driven variance modeling is effective for distinguishing outliers (bad) from healthy parts (good). Variance reduction has helped identify and discard reliability at-risk dies and completely avoid burn-in at the 180-nm and 130-nm technology nodes, and results are in development for the 90-nm and 65-nm technology nodes. The challenge, however, is to identify which of the hundreds of measurements should be postprocessed for variance reduction and used as a screen for burn-in reduction. Adding data mining techniques to the test engineer's arsenal of statistical-test techniques could help address this challenge.

χ^2 response selection

In a production environment, hundreds of tests are performed on each die at wafer probe or package test, but only a subset of those parameters can be leveraged for outlier screening. Data mining is the systematic search of a large volume of unstructured data for subtle patterns that can be extracted and used in downstream applications. Data mining the data can help define a new metascreen, which is a combination of two or more test responses. The metascreen is a quality improvement tool and is an example of statistical prediction (inference). That is, on the basis of combined wafer tests mined from the data, a part is predicted to become a burn-in failure or predicted to fail if inserted into the customer application. The core of the analysis is identification by correlation of a subset of tests that combine to detect the maximum number of burn-in failures while minimizing the test escapes that are not stressed and fail in customer applications.

One of the simplest but most effective techniques of data mining is to hypothetically categorize data and then test whether the hypothetical categories are statistically independent. If they are not independent, a potentially useful correlation exists. A critical step is to choose criteria to do the categorization. Many different ways of doing this are possible for any given data set. In the example here, one of the categories is always passing or failing after burn-in, and another category is chosen to be any of many different wafer sort tests. For any given categorization, it's easy to use the χ^2 test to detect useful dependencies. For example, if data can be categorized in two ways (in one way with a categories, and in another way with b categories), then we can compute the χ^2 statistic by using the data to estimate marginal category probabilities p and q:

$$p_i = \sum_{j=1}^{b} \frac{X_{ij}}{n}, \quad q_j = \sum_{i=1}^{a} \frac{X_{ij}}{n} \tag{12}$$

where n is the number of parts in the data set, and X_{ij} is the count of units in the ith cell of one category and the jth cell of the other. This is a *contingency table*. The probabilities are used to estimate the χ^2 statistic:

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(X_{ij} - np_iq_j)^2}{np_iq_j} \tag{13}$$

In the data mining example, $a = b = 2$. The independence test compares the computed χ^2 value to a critical value set by a statistical confidence limit and the degrees of freedom within the data.[2] When the computed χ^2 is larger than the critical value, the statistical independence of the sort test and burn-in response is not supported by the data.

Identifying a wafer sort test metascreen involves four steps: data set identification, wafer outlier identification, χ^2 response selection, and metascreen validation.

Step 1: Data set identification. Data selection is a simple step. Typically, data mining begins with a selection of a sample of historical data thought to contain the characteristics of future data inputs. The metascreen is product data specific, so results from two data selections do not necessarily agree 100% with each other. For burn-in reduction, the data analyzed was a combination of all wafers from the wafer sort

database containing one or more parts that failed earlier burn-in reliability studies.

Step 2: Wafer outlier identification. Many articles discuss outlier identification, so we will not discuss it in detail here. Once the wafers are selected for analysis, test limits are needed to isolate the outliers. The literature includes many examples of outlier screening.[7] Some standards are emerging for techniques such as *part average testing* (PAT) (http://www.aecouncil.com/AECDocuments.html). These standards include dynamic limits to identify outliers without excessive overkill. In this study, we used the *location averaging* nonparametric technique to compute the response residuals.[6] For every parameter, we obtained a list of outlier parts by setting a confidence bound about a linear regression of each die's location average estimate and its residual. Figure 5 shows an example of this computation.

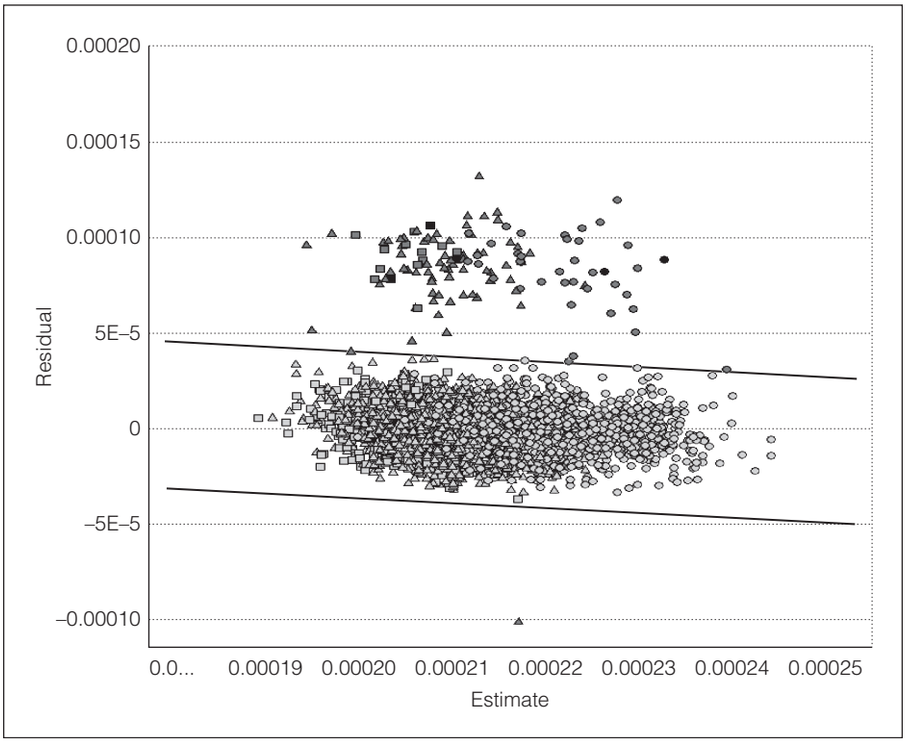

Figure 5. Robust regression of a test response residual versus its estimate. Dies outside the prediction limits (solid lines) are outliers.

Step 3: χ^2 response selection. In this step, we obtain a list of outliers for each test response parameter and wafer. (This step correlates the burn-in failures to the outliers at wafer probe.[8]) Not all burn-in failures will necessarily have a test response in an outlier region such as the one in Figure 5. If a burn-in failure is an outlier to a test, we create a 2×2 contingency table for the RAM0 I_{DDQ} test, as Table 1 shows. (There are two different memories on this chip: RAM0 and RAM1. These memories are in separate regions of the chip, and the size of each memory is different. This example uses burn-in failures, but we can replace this label with *customer returns* or any quality label. For each 2×2 contingency table, we compute the χ^2 and the corresponding p value (the probability that the contingency table is the result of two

Table 1. Example 2×2 contingency table for testing the independence of a test (RAM0 I_{DDQ}) to the post-burn-in test response.

	Outlier sort response to the post-burn-in test					
Post-burn-in test result	**RAM0 I_{DDQ} outliers**	**RAM0 I_{DDQ} nonoutliers**	**Total**	**χ^2 computed value**	**χ^2 critical value**	***p* value**
Burn-in failures	5	12	17			
Burn-in passes	381	26,077	26,458	92	6.63	< 0.0001
Total outlier columns	386	26,089	26,475			

* The confidence level $\alpha = 0.01$, and the degrees of freedom $df = 1$ for this χ^2 test. The p value is the probability that the contingency table is the result of two statistically independent random variables.

Table 2. Summary of three top wafer sort tests shown by χ^2 to most likely correlate to the observed burn-in failures.

Test response	χ^2	*p* value	Detected failures (cumulative %)	Outliers (cumulative %)
RAM0 I_{DDQ}	93.7	< 0001	29	1.4
V_{DIFF}	41.0	< 0001	47	3.0
RAM1 I_{DDQ}	29.1	< 0001	53	3.4

* V_{DIFF} is a parametric test measuring the differential voltage between the input and output of an internal voltage regulator.

statistically independent random variables. Tests with a *p* value less than 0.01 are retained for selection (inference of correlation), and tests with larger *p* values are eliminated from the metascreen (inference of independence). The smaller the *p* value, the less likely the rows and columns are independent. Using the selected tests, we construct a *burn-in cover table* (the smallest combination of tests such that a maximum number of burn-in failures fail at least one test in the combination) to obtain the final, optimized metascreen for the sample population of the detected burn-in failures.

Step 4: Metascreen validation. The final validation involves a review of the metascreen. In practice, the review has exposed remarkably few retained but unrealistic test responses. The data mining flow identifies a wafer sort metascreen of burn-in failures or customer returns. Each parameter is tested by χ^2 as a statistically significant correlation, and the final cover table is minimized. However, we constructed the metascreen from historical data and without reference to the specific post-burn-in tests failed by each part in the sample. When it works, data mining reveals valuable but often subtle relationships in the data. But it is not a replacement for good sense and engineering judgment. A sanity check of each part's observed failure mode to the retained tests in the metascreen builds additional empirical confidence in the metascreen's promise to predict future burn-in failures with only these wafer sort tests. After validation, unrealistic test responses are removed before the metascreen is deployed.

Data mining metascreens

We conducted a case study to demonstrate parametric responses of data mining screening for burn-in reduction and the effects of the outlier limit setting and the 2×2 χ^2 test for independence. We have obtained the results of the data mining test response reduction and metascreen for a broad range of product lines and device technologies (more than 50, thus far). The analysis includes products manufactured across multiple foundries and across multiple internal production facilities. We have extracted metascreens via data mining and deployed them in production to meet goals for burn-in reduction, burn-in elimination, and customer DPPM (defective parts per million) reduction. The case study demonstrates that subtle signals in test data can be used to screen reliability failures as well as customer returns.

The case study used a 65-nm low-leakage device. The goal of the analysis was to identify outlier screens that could predict burn-in failures at wafer probe to avoid the costs associated with burn-in. We applied burn-in stress to a sample population and observed post-burn-in failures. A burn-in cover table minimized the list of wafer sort tests identified by the χ^2 statistic as sort tests significantly correlated to burn-in failures.

Table 2 gives the details of the metascreen. The results are compiled for two related figures of merit for retention (χ^2 and *p* value). The three listed tests each had a less than 0.01% probability of being independent of a part's burn-in failure. Two key figures of merit for the metascreen are the detected failures and the total outliers screened. In Table 2, for each test added to the metascreen, the former represents the total fraction of post-burn-in failures screened; the latter reflects the cumulative cost of using the metascreen, as measured by the total fraction (burn-in failures + burn-in passes) of parts identified by the metascreen. For example, outliers of the RAM0 I_{DDQ} parameter alone could identify 29.4% of the burn-in failures, with 1.4% of all outlier parts screened. The combination of all parameters would identify 53% of the burn-in failures, with 3.4% parts screened. The validation further confirms that metascreen tests show a meaningful relation to the failure modes observed for the burn-in failures.

The role of the final metascreen is not to replay results but to improve future testing. The metascreen can direct a small fraction (estimated to be less than 4%) of the material to burn-in and still detect more than 50% of final stress failures. For this part, any screen identifying at least 50% of the burn-in failures achieved DPPM goals. The metascreen met this goal without burn-in, and 4% of units directed to and passing burn-in were recovered for no loss in yield. By reducing the number of parts stressed at the burn-in metascreen, there is an attendant cost savings

because burn-in reduction reduces the time between the product design and product introduction.

Many researchers outside test can contribute to the foundation of statistical test. The untapped reservoir of statistical techniques applied to the rich data sets in test engineering will provide an interesting field of research for many years to come, as well as opportunities to simultaneously reduce costs and improve quality. ■

Acknowledgments

On the opening day of the 2008 International Test Conference, the organizing committee offered two tutorials of broad interest to the test community. W. Robert Daasch presented one of these tutorials: "Outliers and the Testing Tools That Reveal Them: A Fair and Balanced Introduction to Statistics in Test." This article is a follow-up to that tutorial. We thank Semiconductor Research Corporation for its financial support to Portland State University grant 423425. We also thank the reviewers for their comments and input to improve the article.

References

1. K. Butler et al., "Multi-dimensional Test Escape Rate Modeling," *IEEE Design and Test,* vol. 26, no. 5, 2009, pp. 74-82.
2. R.V. Hogg, A. Craig, and J.W. McKean, *Introduction to Mathematical Statistics,* 6th ed., Prentice Hall, 2005.
3. T. Hastie, R. Tibshriani, and J. Friedman, *Elements of Statistical Learning: Data Mining, Inference, Prediction,* 2nd ed., Springer, 2009.
4. N. Velamatti and W.R. Daasch, "Analytical Model for Multi-site Efficiency with Parallel to Serial Test Times, Yield and Clustering," *Proc. 27th IEEE VLSI Test Symp.* (VTS 09), IEEE CS Press, 2009, pp. 270-275.
5. C.H. Stapper, "Modeling of Integrated Circuit Defect Sensitivities," *IBM J. Research and Development,* Nov. 1983, pp. 549-557.
6. W.R. Daasch et al., "Neighborhood Selection for I_{DDQ} Outlier Screening," *IEEE Design and Test,* vol. 19, no. 5, 2002, pp. 74-81.
7. S. Sabade and D. Walker, "Use of Multiple IDDQ Test Metrics for Outlier Identification," *Proc. 22nd IEEE VLSI Test Symp.* (VTS 03), IEEE CS Press, 2003, pp. 31-38.
8. A. Nahar, R. Daasch, and S. Subramaniam, "Burn-in Reduction Using Principal Component Analysis," *Proc. IEEE Int'l Test Conf.* (ITC 05), IEEE CS Press, 2005, pp. 155-164.

W. Robert Daasch is a professor of electrical and computer engineering at Portland State University. His research interests focus on statistical test for semiconductor testing. He has a PhD in chemistry from the University of Washington, Seattle. He is a member of the IEEE and Sigma Xi.

C. Glenn Shirley is an adjunct research associate professor in the Integrated Circuits Design and Test Laboratory at Portland State University. His research interests include test manufacturing, packaging, burn-in, and reliability modeling. He has a PhD in physics from Arizona State University.

Amit Nahar is a test engineer at Texas Instruments. His research interests include outlier techniques for quality and reliability, and adaptive test solutions based on test data. He has an MS in electrical and computer engineering from Portland State University.

■ Direct questions and comments about this article to W. Robert Daasch, Integrated Circuits Design and Test Laboratory, Electrical and Computer Engineering, Portland State University, Portland, OR 97207; daasch@ece.pdx.edu.