An Innovative Method to Hypothesis Testing for System Safety Assessment

by Dr. Robert W. L. Thomas, Marilyn J. Eichelberger, Missey Lee and Joel Haan
Bowie, Maryland and Dahlgren, Virginia

The way forward in system safety engineering will be quantitative, and this paper proposes an innovative method for generating a uniform way to understand the composite of testing and experience. In recent years, new approaches to exact hypothesis testing have been developed without a Gaussian probability distribution for success or failure rates. These techniques eliminate errors introduced by the Gaussian assumption, which is important for the small failure rates that are common in modern systems development, and offer considerable promise as a basis for the new direction.

This paper presents a theory for exact hypothesis testing and combines two 18th-century theorems to derive an equation for the probability distribution of failure rate employing only the number of tests and the observed count of failures. The concept is expanded to demonstrate the combination of operational experience and expert opinion to update test results. The objective in this work is to derive the general likelihood distribution of failure rate given any set of test results, and then to examine the implications regarding testing requirements, design and interpretation. The particular application considered here is safety assessment for a military weapons system. While the theory developed is for deriving the exact failure rate distribution for system safety applications, it is equally valid for investigating success rates and/or for interpreting performance evaluation tests.

Introduction

In recent years, new approaches to exact hypothesis testing have been developed — approaches that do not depend on the assumption of a Gaussian probability distribution for the success or failure rates [Refs. 1-3]. These approaches are helpful in eliminating errors introduced by the Gaussian assumption, particularly for the small failure rates that are common in modern systems development. While the theory developed here is for deriving the exact failure rate distribution for system safety applications, it is equally valid for investigating success rates and/or for performance assessment tests.

This approach provides critical support for Quantitative Risk Analysis (QRA), which is becoming a key component of modern risk mitigation methods [Refs. 4 and 5]. We begin by presenting the theoretical basis for exact hypothesis testing and derive an equation for the probability distribution of the failure rate employing only the number of tests and the observed count of failures. The theory is expanded to show how we combine the composite of a priori data together with expert opinion and updated test results.

Given the importance of increasing the test number when the desired probability is low, we examine results from a typical test matrix designed principally to demonstrate system performance, but with embedded safety statistics included. Limitations will be noted and suggestions made as to how to improve the comprehensiveness of the testing program regarding safety issues.

Motivation

With complex systems, there may be a tendency to believe that a small change may not cause any new safety issues. In this case, there may be no test, or only a single test, for safety flags. Anecdotal evidence suggests that sometimes tests are repeated until an instance is found with no safety problem. The difficulty is that with limited or no understanding of the original system design, it may be difficult to assess the safety impact of a change without extensive testing.

Our objective in this work is to derive the general likelihood distribution of the failure rate given any set of test results, and then to examine the implications regarding testing requirements, design and interpretation. We begin by making the assumption of no prior data. Suppose we have no test and no source of information about the safety of the operation of the system. All we can say is that the failure rate is a probability, uniformly distributed between zero and one. Thus, the average failure probability in this case is one-half. Now consider making a single safety test. If the system passes this test, then we will show that the most likely failure rate is 0, but the average rate is 33 percent. If the system fails the test, the most likely failure rate is 1, but the average rate is now 66 percent. In general, if a total of N tests are passes with no failures, the most likely failure rate is zero, but the average failure rate is 1/(N+2). If all but one fails, the most likely failure rate

is 1 - 1/N, and the average rate is N/(N+2).

Analysis

The situation that we want to describe is one in which a given number, N, of tests are performed (usually by computer simulation) and we observe safety issues in n cases. If the failure probability is p, then the probability of obtaining n failures is given by the Bernoulli formula [Ref. 6]:

$$p(n|p) = [N!/(n!(N-n)!)]p^n(1-p)^{N-n} \quad (1)$$

What we want to know is the probability density, $\pi(p|n)$; i.e., the probability distribution of p, given an observed value of n failures in N tests. Fortunately, this can be done using Bayes' theorem [Refs. 7 and 8], which connects the probability of occurrence of an event, A, with the truth of a hypothesis, B, governing its behavior:

$$P(A|B) = P(B|A).P(A)/P(B) \quad (2)$$

where P(A) and P(B) are the *a priori* probabilities of A and B, respectively.

Since there are (N+1) possible values for n, we obtain from the Bayes' formula:

$$\pi(p|n) = p(n|p).p(p)/p(n)$$
$$= (N+1)p(n|p) \quad (3)$$
$$= [(N+1)!/(n!(N-n)!)]p^n(1-p)^{N-n}$$

Note that $\int_0^1 \pi(p|n)dp = 1$, as expected, and that the cumulative distribution function,

$$\varphi(x) = \int_0^x \pi(p|n)dp \quad (4)$$

relates to the incomplete beta function.

The moments of $\pi(p|n)$ can readily be evaluated using gamma functions of integer arguments as:

$$E(p^k) = (n+k)!(N+1)!/[n!(N+k+1)!] \quad (5)$$

and computed using the recursion relationships:

$$(N+k+1)E(p^k) = (n+k)E(p^{k-1}) \text{ with } E(p^0) = 1 \quad (6)$$

The most likely value of p maximizes $\pi(p|n)$ and is, simply, n/N. On the other hand, the mean value of p, computed from equation 5 by setting k to 1, is:

$$\langle p \rangle = (n+1)/(N+2) \quad (7)$$

Note that, for a single test (N=1), the average failure rate is 33 percent for a pass and 67 percent for a fail, as discussed earlier. The variance of p comes out to be:

$$V(p) = (n+1)(N-n+1)/[(N+2)^2(N+3)] \quad (8)$$

Application Considerations

We now show how the functions given above are required to develop a simulator of accident occurrence rates. They are readily available in most mathematical or spreadsheet programs. For example, in Microsoft Excel, p(n|p) and $\pi(p|n)$ can be calculated using the BINOMDIST function, while φ can be estimated using BETADIST. Specifically, the functions should be used with arguments as follows:

$$p(p|n) = \text{BINOMDIST}(n, N, p, \text{FALSE}) \quad (9)$$

$$\pi(p|n) = (N+1)*p(p|n) \quad (10)$$

$$\varphi(x) = \text{BETADIST}(x, n+1, N-n+1) \quad (11)$$

The argument, "FALSE," in equation 9 indicates that the terms should not be summed from 0 to n.

Equation 11 is useful for generating a random simulated failure probability with the correct distribution. We simply solve for x when $\varphi(x)$ is equal to a uniformly distributed random number, ρ, between 0 and 1; i.e., x is the solution of

$$\varphi(x) = \rho \quad (12)$$

This is conveniently accomplished using the Microsoft Excel function, BETAINV, as follows:

$$x = \text{BETAINV}(\rho, n+1, N-n+1) \quad (13)$$

It should be noted that even when n is zero — i.e., no failures are observed in the tests — a non-zero failure probability will always be generated by this procedure.

The final step in the simulator is to test to see if a new uniformly distributed random number, ρ, between 0 and 1 is less than x. If so, a failure is declared, but otherwise, the fault does not occur. This step is applied to all bottom-level events in the fault tree, and the results are combined according to the logic of the

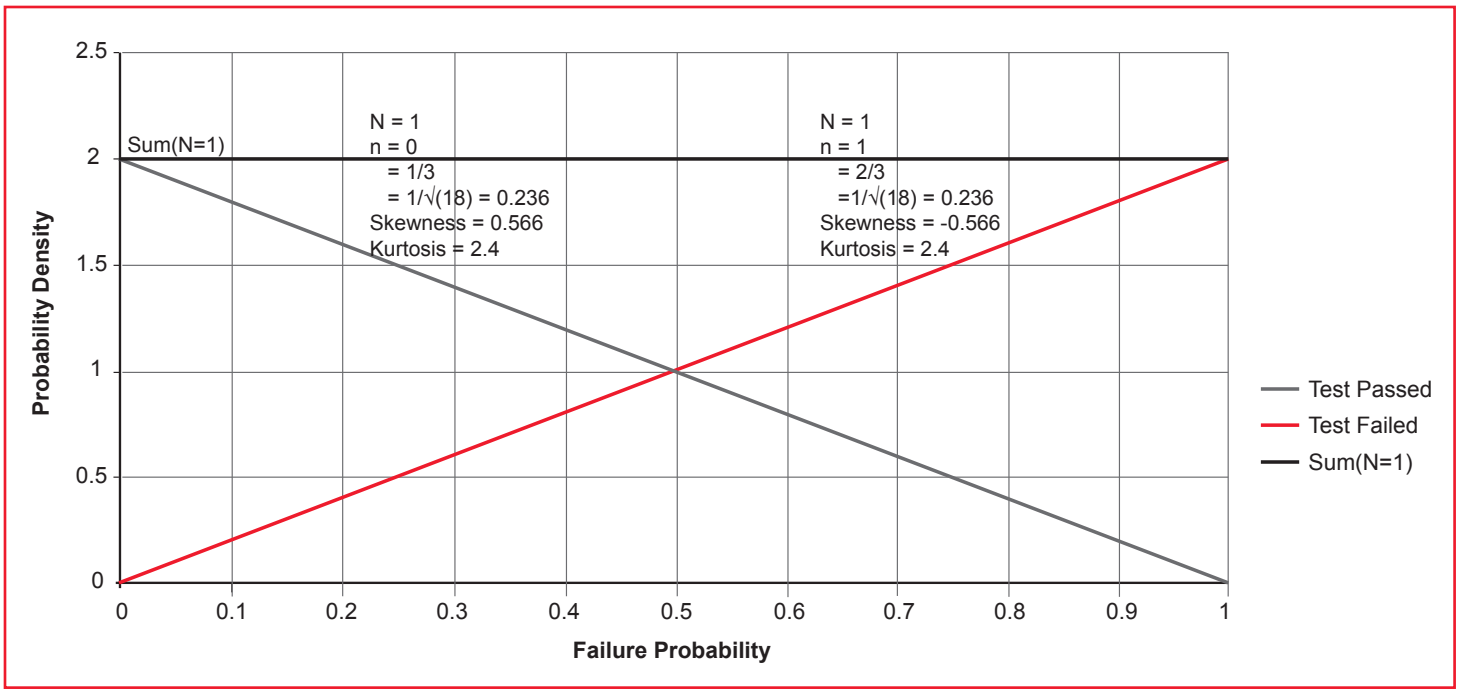

Figure 1 — Probability Distributions for a Single Test.

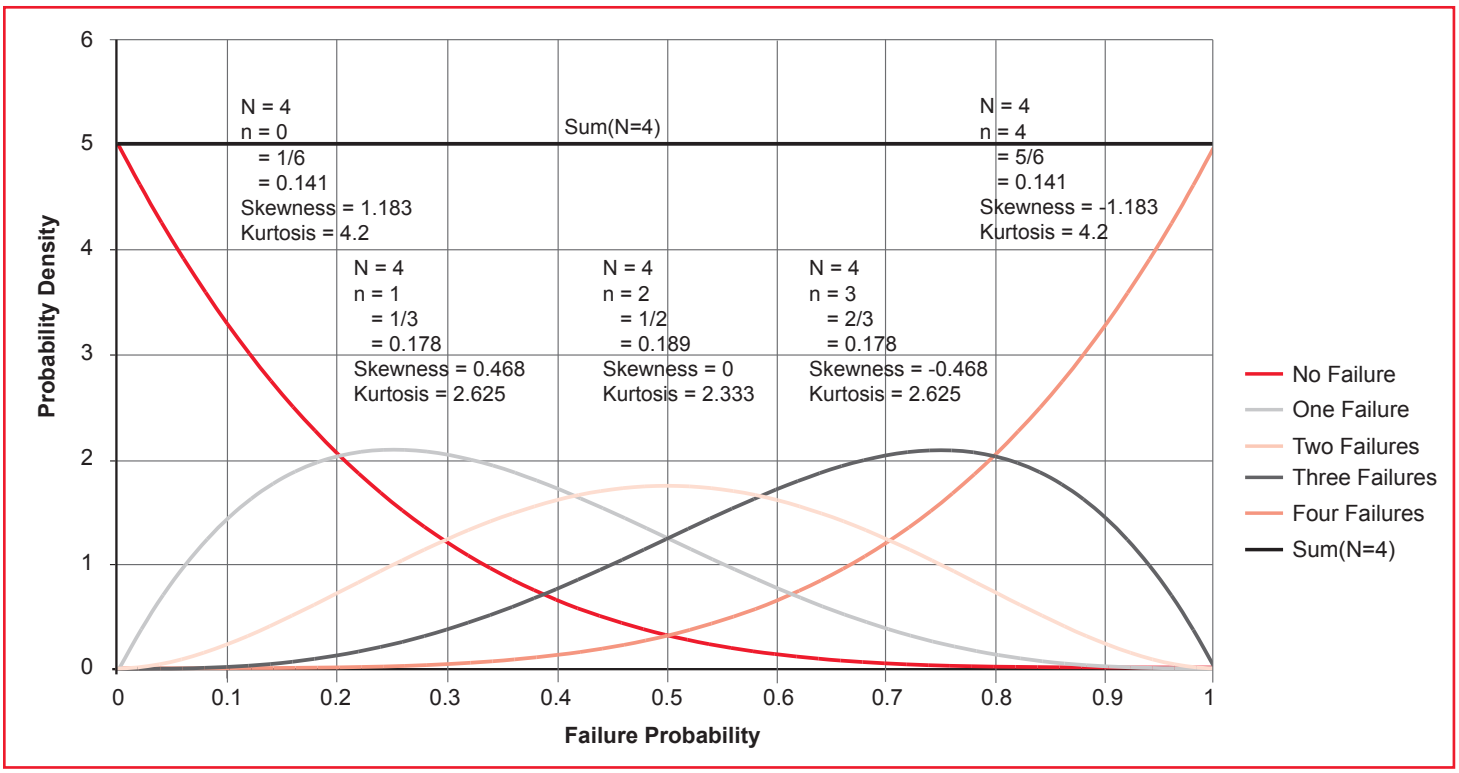

Figure 2 — Probability Distributions for Four Tests.

fault tree. If there are any Common Causes of Failure (CCFs), then each CCF needs to be simulated only once and the results copied to all relevant nodes of the tree.

Figures 1 and 2 illustrate some aspects of the predicted distribution, $\pi(p|n)$, of the failure rate, p, given n test failures. Figure 1 shows the computed distribution given zero and one failure of a single test, while Figure 2 shows the results with four tests. In all cases, the sums of the probability densities are equal to N+1. The parameters listed on the graphs are the mean failure probability, μ; the standard deviation, σ; the skewness and the kurtosis. The former indicates the degree and direction of the asymmetry, while the latter is an indi-

Table 1 — Fault Tree Statistics for the Aegis Weapons System.

Top-Level Mishaps	Numbers of Intermediate Events	Numbers of Basic Events
TLM 1	8	107
TLM 2	5	40
TLM 3	3	18
TLM 4	8[1]	24[2]
TLM 5	7[1]	26[2]
TLM 6	1	1

[1] Four CCFs were identified
[2] Six CBEs were identified

cation of how peaked the distribution is. The skewness of a Normal distribution is zero, while the kurtosis is 3, so the distributions, $\pi(p|n)$, are clearly non-Normal.

Certain symmetries in these graphs are useful. Note that $\pi(p|N,n) = \pi(1-p|N,N-n)$, showing that the probability distribution of the failure rate is that of the complement of the success rate distribution.

Table 1 enumerates the Top-Level Mishaps (TLMs) for the Aegis Weapons System (AWS), together with the number of faults identified as causal factors. The fault numbers indicate the number of possible mechanisms by which each TLM can occur. Also listed is the number of basic events developed under each TLM. Note that no further development of TLM 6 was analyzed here. TLMs 4 and 5 had four CCFs, so the simulation result of whether that fault occurred in TLM 4 was also applied in TLM 5. The four CCFs identified at the intermediate level resolved into six Common Basic Events (CBEs), which was the most detailed level used at the lowest level of the fault tree.

If S_i is the average probability of avoiding the occurrence of the ith of N independent terminal nodes of a simple fault tree as we have above, then the overall success probability of the entire system is:

$$S = \prod_{i=1}^{N} S_i \qquad (14)$$

where the product is taken over all independent failure causes. Thus, each CBE is included only once in the calculation. S_i is best computed as the complement of the average failure probability, which can be derived from equation 7. The result is:

$$S_i = 1 - \langle p_i \rangle = 1 - (n_i + 1)/(N_i + 2) = (N_i - n_i + 1)/(N_i + 2) \qquad (15)$$

where n_i and N_i are the observed failure number and number of tests, respectively, for the ith terminal node. For example, if two failures were observed in 100 tests, the predicted average success rate would be 99/102. Note that even if no failures at all were observed in any of the tests, the average predicted success rate would still not be unity; i.e., if no failures were observed in 100 tests, the average success rate is only 101/102 and not 102/102. Hence, it is clear that absent any other data than the tests, the number of independent tests is vital.

Figure 3 presents the results of computed system success rates (i.e., no safety incidents) in a simulation of the AWS, with the assumption that 100 tests were performed on each CCF. Figure 4 presents an analytical approximation using the mean success rate given by equation 15 as input to equation 14. It can be seen that the agreement is excellent. It must be emphasized that the modeled failure rate is given by equation 7; i.e., the mean probability and not the most likely value.

Incremental Assessments

Equation 3 gives a distribution of the beta form. It has been shown that if the a priori distribution is of this form, then all subsequent distributions derived from this will also be of the beta form [Ref. 9]. An important question to address is the use of experience and expert opinion gained in addition to the testing that is done. This knowledge can be gained from past experience with legacy systems that have been modified only in minor ways. In our case, an appropriate procedure is to weight prior estimates of the failure rate using an assumed number of equivalent tests. For example, if prior testing indicated a homogeneous failure rate of about 2 percent, then we might estimate that this was derived from 200 prior tests, with four observed failures. In this case, we would add 200 to the value of N, and 4 to the value of n used in equations 1 through 15. Obviously, if the number of equivalent tests in the prior experience is large, then a larger number of new tests will be required to significantly change the prior distribution estimate.

Conclusions

While it is impossible to derive the failure rate by testing, it is possible to derive the exact probability distribution of the rate, given the test results. Thus, the test data can always be interpreted as probabilities

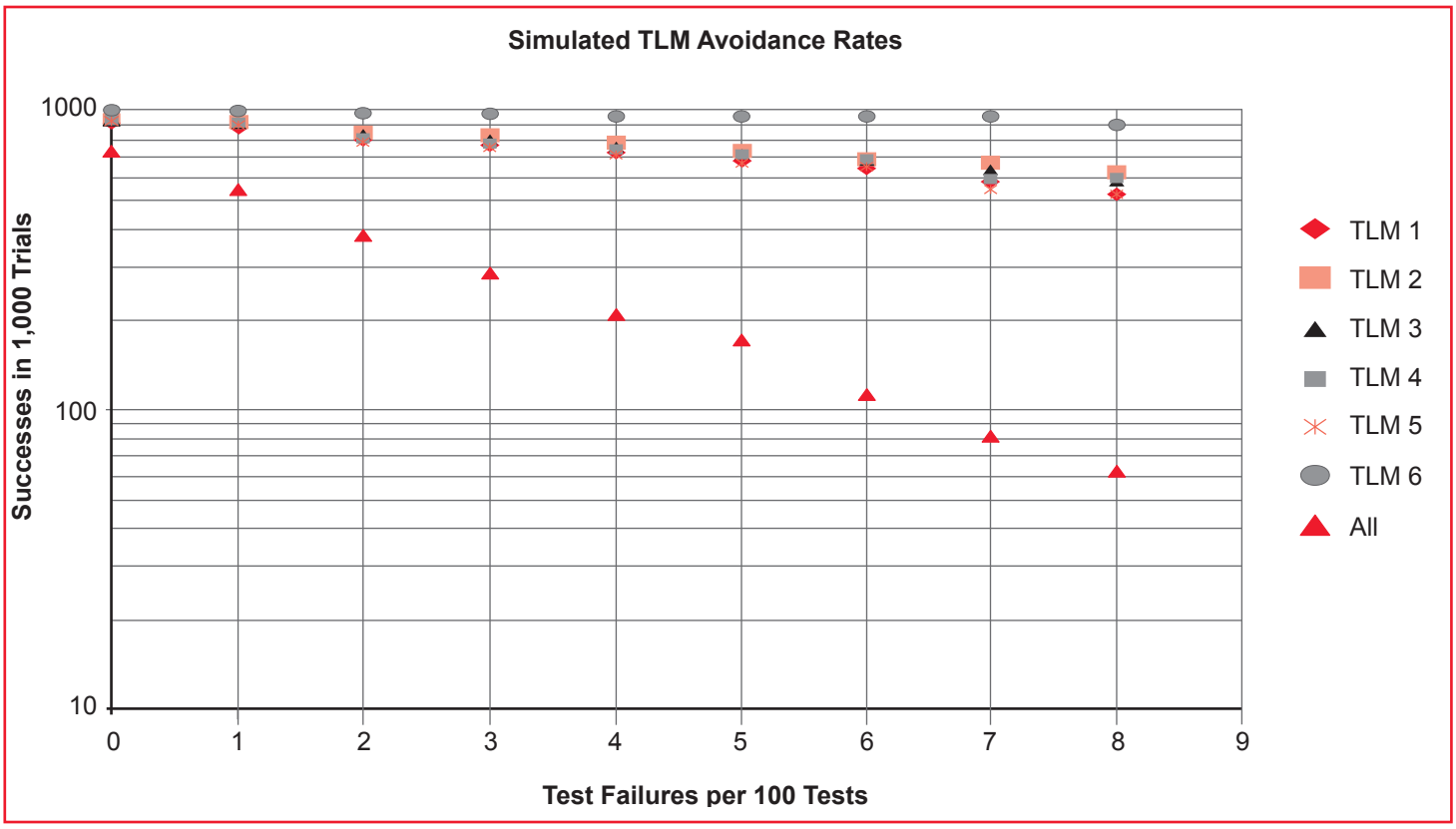

Figure 3 — Simulated Number of Cases with Zero Failures in 1,000 Trials.

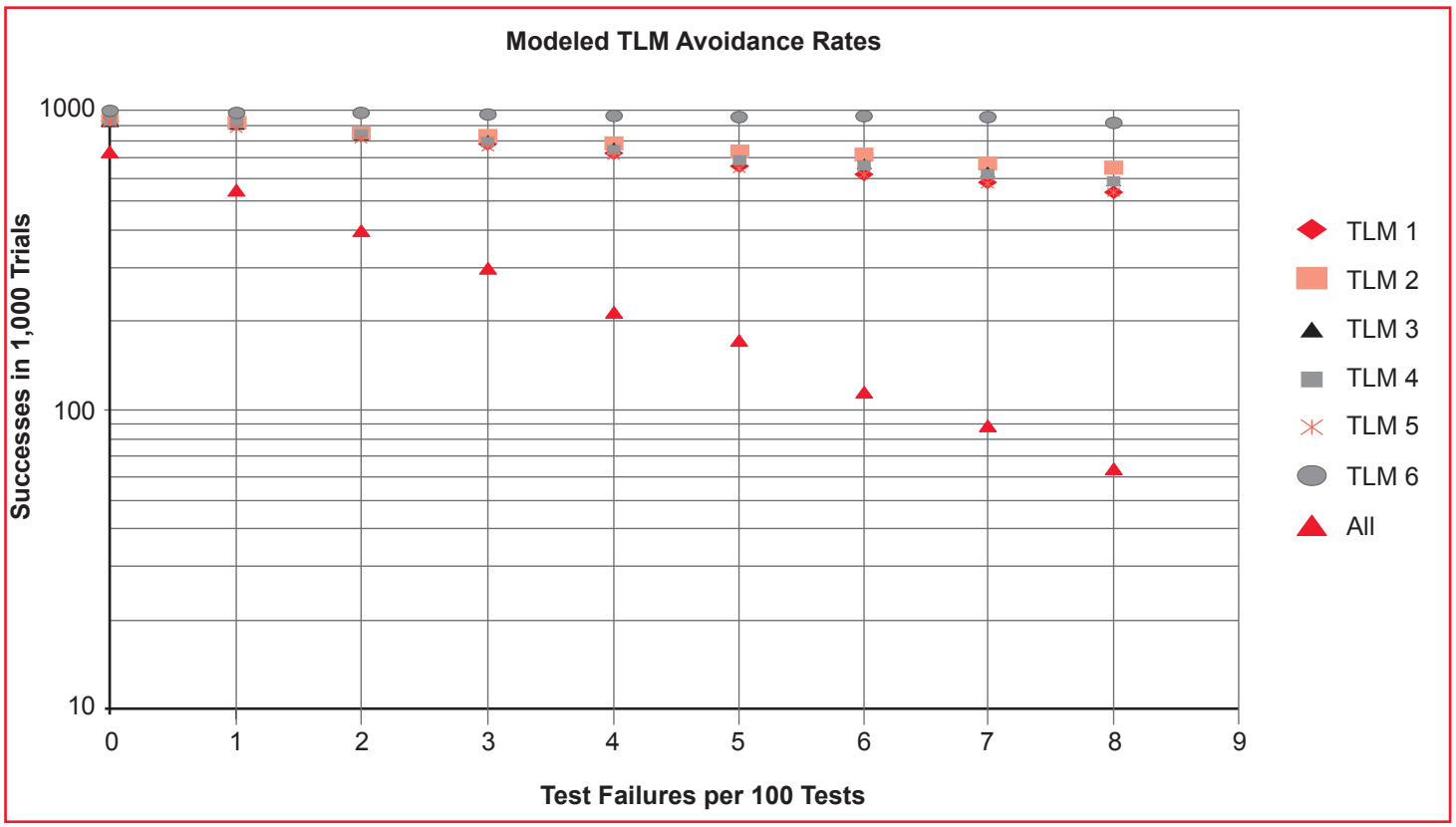

Figure 4 — Modeled Number of Cases with Zero Failures in 1,000 Trials.

that the failure rate is less than any set of values. We have given a procedure for making simple estimates of these probabilities.

For the particular system treated, the test numbers sometimes had to be derived indirectly. Nevertheless, it was always possible to represent the test data as the number of successes or failures for a specified number of tests. The derived test counts were often small, or zero, suggesting that a much larger level of testing might be necessary to ensure the safety levels required. Expert opinion is represented as equivalent prior testing experience, providing a uniform rational method for appropriate weighting.

We show how the methods can be applied to performance testing, as well as to safety testing. The comments concerning the increased level of testing required to raise safety confidence are equally applicable to performance test requirements. Future effort will focus on expanding the interpretation of these methods with prior and future tests.

About the Authors

Dr. Robert W. L. Thomas is a scientist at AECOM working on safety issues pertaining to electromagnetic environments near to high-power emitters. He has experience with the verification and validation of combat systems software, with a particular emphasis on critical timing issues, and has also developed models used by the National Aeronautics and Space Administration (NASA) to scope the independent verification and validation of satellite and spacecraft-borne software. He has also designed optical instrumentation for space measurements to characterize planetary atmospheres.

Marilyn J. Eichelberger is currently a member of the Combat System Safety Branch at the Naval Surface Warfare Center, Dahlgren Division. She has 20 years of experience supporting U.S. Navy, U.S. Marine Corps and foreign military programs, of which seven years have been in the role of Principal for Safety.

Missey Lee is currently a member of the Platform System Safety Branch at the Naval Surface Warfare Center, Dahlgren Division. She has more than 10 years of experience practicing system safety engineering for the U.S. Navy. Prior to this, she worked on developing and implementing code for weapon systems of the U.S. Navy.

Joel Haan is currently a member of the Engagement System Safety Branch at the Naval Surface Warfare Center, Dahlgren Division. He has five years of experience practicing system safety engineering for U.S. Navy and U.S. Marine Corps safety programs. Prior to this, he worked on test engineering for both developmental and in-service weapon systems.

References

1. Schlag, K.H. "Exact Hypothesis Testing without Assumptions — New and Old Results not only for Experimental Game Theory" http://homepage.univie.ac.at/karl.schlag/research/statistics/exacthypothesistesting8.pdf, October 11, 2013.
2. Schlag, K.H. "Exact Hypothesis Tests for Experimental Game Theory," https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=WCGTS2007&paper_id=186, accessed June 5, 2013.
3. Ferson, S. "Bayesian Methods in Risk Assessment" (2012), http://www.ramas.com/bayes.pdf, accessed June 5, 2013.
4. Vose, David. *Risk Analysis: A Quantitative Assessment*, John Wiley and Sons, Ltd., (2008), http://books.google.com/books?hl=en&lr=&id=9CaoAqaRcVwC&oi=fnd&pg=PR13&dq=quantitative+risk+analysis&ots=tOfXfEsqT1&sig=IV3Sg57WkRDxfYPGUKq0yVdP9vc, accessed June 5, 2013.
5. Sims, Stephen. "Qualitative vs. Quantitative Risk Assessment" http://www.sans.edu/research/leadership-laboratory/article/risk-assessment, accessed June 5, 2013.
6. Simmons, Bruce. "Bernoulli Probability Formula," http://www.mathwords.com/b/binomial_probability_formula.htm, accessed June 5, 2013.
7. Joyce, James. "Bayes' Theorem," *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/fall2008/entries/bayes-theorem/, accessed June 5, 2013.
8. Yudkowsky, Eliezer S. "An Intuitive Explanation of Bayes' Theorem" http://yudkowsky.net/rational/bayes, accessed June 5, 2013.
9. Navarro, Daniel and Amy Perfors. "An Introduction to the Beta-Binomial Model" http://www.cs.cmu.edu/~10701/lecture/technote2_betabinomial.pdf.