

Approaches to Achieving High Reliability and Confidence Levels with Small Test Sample Sizes

Zhigang Wei, Robert Rebandt, Michael Start, Litang Gao, Jason Hamilton, and Limin Luo
Tenneco Inc.

ABSTRACT

In product design and development stage, validation assessment methods that can provide very high reliability and confidence levels are becoming highly demanded. High reliability and confidence can generally be achieved and demonstrated by conducting a large number of tests with the traditional approaches. However, budget constraints, test timing, and many other factors significantly limit test sample sizes. How to achieve high reliability and confidence levels with limited sample sizes is of significant importance in engineering applications. In this paper, such approaches are developed for two fundamental and widely used methods, i.e. the test-to-failure method and the Binomial test method. The concept of RxxCyy (e.g. R90C90 indicates 90% in reliability and 90% in confidence) is used as a criterion to measure the reliability and confidence in both the test-to-failure and the Binomial test methods. For the test-to-failure method, high reliability and confidence levels are achieved by using historical data and the Bayesian statistics approach. For the Binomial test method, the high levels are achieved through extending test time and accelerating tests by increasing the stress/load level. With these methods, high reliability and confidence levels, such as R99C99 or higher, can be achieved with small sample sizes.

CITATION: Wei, Z., Rebandt, R., Start, M., Gao, L. et al., "Approaches to Achieving High Reliability and Confidence Levels with Small Test Sample Sizes," *SAE Int. J. Commer. Veh.* 8(2):2015, doi:10.4271/2015-01-2758.

1. INTRODUCTION

Durability and reliability performance of products is a major concern for many engineering components and systems [1, 2, 3]. Numerous methods are available for product durability and reliability demonstration, and such methods include the test-to-failure method and the Binomial test method [4, 5, 6, 7, 8]. The test-to-failure method tests a component to the occurrence of failure under the specified loading. The Binomial test method tests the product to the target under specified loading conditions. Only two possibilities occur in the Binomial test: pass or fail. Test to success, also referred to as success-run method, is a special Binomial test method, in which no failure occurs. The extended test method, which is an extension of the Binomial test method, test a component or a system for a time longer (usually <1.5X) than the bogey under the prescribed loading [5, 6, 7]. The extended testing method trades additional test time for smaller sample size requirements, provided that the required parameters, such as the Weibull slope, of the life distribution are given. The test-to-failure method provides the life distribution for a product, but the test-time-to-failure is usually longer. By contrast, the Binomial test method often requires shorter test time but a relatively larger sample size for the same reliability and confidence levels. It should be noted that the original Binomial test method and the extended test method cannot reveal the life distribution of a product [4, 5, 6, 7].

Test data with failure, such as fatigue test data, usually show significant scatter and, therefore, repeated tests are often required to capture the uncertainty, scatter, and failure probability of the test data.

Based on test data, several statistical parameters, including mean and standard deviation, can be estimated. However, the mean and standard deviation are usually not of primary interest. In fact, engineers, particularly durability and reliability engineers are more interested in specific measures of product reliability or particular failure probabilities [5, 8]. One such measure is design limit when one-stress level testing is conducted or design curve when two-stress level testing is performed [9].

For test-to-failure method, a design limit or curve is generally constructed to guarantee that a majority of the data fall above the lower bound with certain failure probability and confidence levels. The common practice to define a design limit or curve is to shift the mean point or curve of the failure data by a distance to the left of the mean in order to achieve a certain "Reliability" level [9, 10, 11, 12, 13, 14, 15]. The traditional methods [14, 15] require significant amounts of test data to reduce the uncertainty in estimating the population behavior, but at the expense of test cost and time. Modern design curve construction methods are based on test data with limited sample size, which usually requires advanced statistical analysis tools to properly handle the uncertainty introduced by the sample size effect [9, 10, 11, 12, 13]. The tolerance limit and probability plot methods are such two typical approaches, in which the uncertainty caused by sample size effect is quantitatively described using the concept of "Confidence". Design criteria of R90C90 (i.e. 90% reliability and 90% confidence) and R99C50 are commonly used in the automotive industry for fatigue data analysis [1]. Six samples at

each stress level for two stress levels, therefore, 12 samples are commonly used for constructing a linear (in a log-log plot) design curve in bench test for vehicle product validation [9].

Although these test methods and procedures are routinely used in engineering design and validation, new approaches, which can reduce sample size and associated test cost/time and increase assessment accuracy, are also intensely investigated. Bayesian statistics is such a method, and it generates posteriori distribution by updating prior distribution based on the historical data and new data input [16]. Recently, several Bayesian statistics based design curve construction methods have been developed, and the goal of sample size reduction has been successfully achieved [17]. In a recent study [18], the probabilistic distributions of the mean and the standard deviation of the failure cycles have been extracted from a large amount of historical fatigue test data of welded structures. These prior distributions together with the currently obtained test data with small sample sizes are used to generate posterior distributions with the help of an advanced acceptance-rejection resampling algorithm [19, 20]. The posterior distributions can subsequently be processed with conventional statistical tools to generate more useful information, such as design limits and curves. In the earlier approach [18], called Bayesian-V1 here, finite element analysis (FEA) methods and a special stress recovery technique, i.e. the Verity method [20], are used to comply with the master S-N curve for welded structures [20]. In a recent paper [21] a simplified Bayesian statistics method, called Bayesian-V2 here, has been developed. Neither FEA nor stress recovery technique is required in this method. The only historical information required is the distribution of the standard deviation of the cycles to failure, which are already available for the type of welded structures [18]. In Bayesian-V2, the distribution of mean is calculated from the current test data input and the Central Limit Theorem, which could make the new method more accurate and reliable than Bayesian-V1 [21]. With these Bayesian statistics approaches, a design limit or curve can be constructed with sample size of 3, 2, and even 1, which is much smaller than the current 6 samples at each stress/load level, eventually leading to achievement of the cost saving goal.

Generally, a design criterion should be a balanced outcome between durability/reliability performance and cost. Usually, a conservative design is preferred in practice; however, over-conservative design should be avoided in order to reduce cost. This is especially important for non-mission-critical engineering such as automotive industry, in which fuel economy, weight reduction, and cost effectiveness etc. are important considerations. With these criteria, it is expected that a majority of the products designed with a "Pass" would be safe in the service life and a majority of the products designed with a "Fail" would show some risks in the service life. However, with this approach, it is not unexpected that some qualified designs can be rejected and some unqualified designs can be accepted due to the uncertain nature of the test data and the simplified single criterion: RxxCyy.

It should be emphasized here that the establishment of the values of xx and yy in RxxCyy criteria such as R90C90 and R99C50 are essentially based on a large amount of data correlation at a specific sample size, say 6, engineering judgment, and engineering verification made over decades. The value of RxxCyy based on the Frequentist (based on

current test data) approach is sensitive to sample size, especially when sample size is very small [9]. For example, if a design limit or curve was constructed based on R90C90 with smaller sample sizes, e.g. 3, the design curve would be over-conservative. Furthermore, with the Frequentist methods, a design curve usually cannot be effectively constructed with sample sizes 2 and 1 [9].

The R90C90 or R99C50 criteria are essentially empirical which is reflected in the fact that a product is usually designed as a few parts per million (PPM) in failure. In other words, the products should be designed with Reliability (R) around 99.999%, which is far higher than R90 and R99. The gap can be partially filled with the introduction of confidence level C with an appropriate value, e.g. 90% or 50%. Calculations involving reliability demonstration are usually based on lower confidence bounds, therefore, by demonstrating 99% reliability at a lower confidence bound we are in fact demonstrating reliability >99% [5]. The current reliability analysis tools are based on certain probabilistic distributions, such as lognormal or Weibull distributions. However, all of these distribution functions are not accurate at tails, therefore, higher reliability, say 99.9% reliability, is usually meaningless with a probabilistic distribution function. With advanced statistical methods, such as the Bayesian statistics method, more accurate and consistent RxxCyy values can be calculated. However, the new design limits and curves are usually on the right side of those obtained using the conventional methods. Therefore, to meet the original reliability goals, say PPM, a new RxxCyy criterion needs be established for the new methods.

For a success-run test, no failure data is available. For a Binomial test with or without failure data the distribution of the failure data is often unknown. However, based on classical statistics theories, a RxxCyy criterion can also be constructed for the Binomial test method [5]. Actually, a recent paper [23] demonstrates the close relationship between the RxxCyy constructed using the test-to-failure method and the RxxCyy obtained using the Binominal test method.

In this paper, the fundamentals of the RxxCyy criterion used in both the test-to-failure method and the Binomial test method are reviewed and thoroughly investigated. For the test-to-failure method, the current Frequentist approach and the Bayesian statistical approach are compared. The need of a new statistically meaningful criterion with high reliability and confidence levels is demonstrated for the Bayesian statistical approach. A procedure to achieve high reliability and confidence levels is also suggested for the Binominal test method.

2. PROBABILISTIC FAILURE AND DAMAGE DISTRIBUTIONS

2.1. Continuous Failure and Damage Distributions for Population

Durability describes the ability of a product to sustain required performance over a period of time or cycles without undesirable failure. Reliability is defined as the probability that a component or a system will function at the customer's expectation level under specific environmental and duty cycle conditions [2, 5]. Durability and reliability test data of components and systems such as cycles to failure from a fatigue test often show large scatter which comes from

many sources, such as material uncertainty, loading uncertainty, and other uncertainties. In many cases, especially for single failure modes, continuous probability density functions (PDFs) $f(N)$ and cumulative distribution functions (CDFs) $F = \int f(N)dN$ can be effectively used for characterizing the scattered test data. For failure with two or more failure mechanisms and modes, the data can be described using bi-modal or multiple-modal distributions [24]. Even though the underlying driving forces of the probabilistic distributions are still not fully understood, the empirical probabilistic analysis approaches play the central role in durability and reliability analyses and product validation.

Fig. 1 [23] schematically illustrates the damage-cycle ($D - N$) diagram for a population. When $N = 0$, the damage $D = 0$; $N = N_f$ indicates a complete failure and $D = 1$. The continuous PDF and CDF can be described as $f[N_f(D=1)]$ and $F[N_f(D=1)]$ provided that the population is already known, e.g. by approximately analyzing the test data with a large amount of samples, say, 100 or 1000. Based on the distribution of the population, a reliability criterion or metric can be established and the robustness of the product design can be evaluated. The most appropriate reliability metric is the simple reliability function $R = 1 - F$: Rxx. For example, a product specification R99 states that 99% of the product is expected to pass a specified target, say, a million cycles. In other words, 1% of the product will fail, i.e. $F = 0.01$, before the million cycle target. Parts per million (PPM) can also be used as a reliability metric, although it is more often used to measure manufacturing quality in production rather than in the design stage [5].

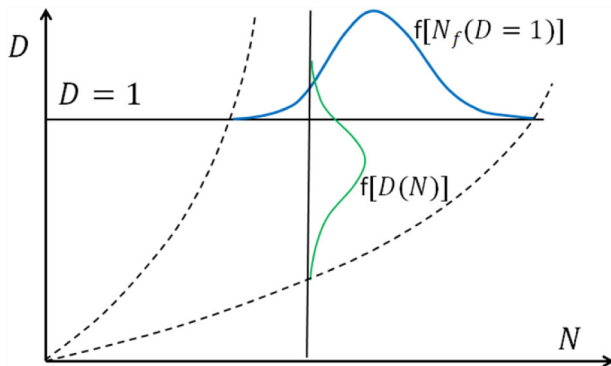


Fig. 1. Schematic of the continuous probabilistic failure distribution $f[N_f(D=1)]$ and the corresponding probabilistic damage distribution $f[D(N)]$ for a population [23].

With the PDF $f[N_f(D=1)]$ and CDF $F[N_f(D=1)]$, the probabilistic distributions of damage at any cycle N (or time t) $f[D(N)]$ and $F[D(N)]$ can be derived provided the damage evolution process is known [25]. Both linear and nonlinear damage evolutions can be used, however, the most commonly used damage law is the linear damage rule or Miner's rule [26]. For the linear damage rule, damage vs. cycle plot is simply a cluster of straight lines. For constant amplitude loading, $D = N/N_f$ and the damages made at each cycle are the same: $D_i = 1/N_f$; i here represents the i th cycle. Mathematically, the problem is equivalent to seeking a target distribution function $F_Y(y)$ for a given initial distribution function $F_X(x)$ and transformation functions: $y = \varphi(x)$ and $x = \psi(y)$. Closed-form solutions can be obtained as follows [27].

$$f_Y(y) = \frac{dF_Y(y)}{dy} = f_X[\psi(y)] \left| \frac{d\psi(y)}{dy} \right| \quad (1)$$

Actually, the two approaches to interpret the life, i.e. $f[N_f(D=1)]$ and $F[N_f(D=1)]$, and damage, i.e. $f[D(N)]$ and $F[D(N)]$, represent the two commonly used testing methods: the test-to-failure and Binomial test methods, which will be demonstrated in next section.

Two examples are provided below to demonstrate how the distribution of damage at any cycle can be obtained by transforming the life distribution using Eq. (1).

Example-1: two-parameter Weibull distribution: For a two-parameter Weibull distribution $f[N_f(D=1)]$ and $F[N_f(D=1)]$, as shown in Eq. (2)₁, and Eq. (2)₂, with the help of the linear damage rule [26], the distribution of $f[D(N)]$ and $F[D(N)]$ can be derived, Eq. (3)₁, and Eq. (3)₂.

$$f(N_f) = \left(\frac{\beta}{\eta} \right) \left(\frac{N_f}{\eta} \right)^{(\beta-1)} \exp \left[- \left(\frac{N_f}{\eta} \right)^\beta \right], \quad N_f \geq 0 \quad (2)_1$$

$$F(N_f) = 1 - \exp \left[- \left(\frac{N_f}{\eta} \right)^\beta \right], \quad N_f \geq 0 \quad (2)_2$$

$$f(D) = \left(\frac{\beta}{\eta} \right) \left(\frac{N}{D^2} \right) \left[\frac{N}{D\eta} \right]^{(\beta-1)} \exp \left\{ - \left[\frac{N}{D\eta} \right]^\beta \right\}, \quad D \geq 0 \quad (3)_1$$

$$F(D) = 1 - \exp \left\{ - \left[\frac{N}{D\eta} \right]^\beta \right\}, \quad D \geq 0 \quad (3)_2$$

where β and η are, respectively, the shape parameter and scale parameter (characteristic life) in the two-parameter Weibull distribution. With the resulting $f[D(N)]$ and $F[D(N)]$, the reliability can be interpreted in a different way, i.e. the possibility of the event at a specific damage value D , in particular, $D = 1$. For example, a product specification may state that the failure possibility of a product at 10-year life should be 1%, in other words, the reliability $R(10 \text{ years}) = 0.99$.

Example-2: Uniform distribution. The PDF and CDF of the uniform distribution function are given in Eq.(4)₁ and Eq.(4)₂, respectively.

$$f(x) = \begin{cases} \frac{1}{b-a}; & a \leq x \leq b \\ 0; & x < a \quad \text{or} \quad x > b \end{cases} \quad (4)_1$$

$$F(x) = \begin{cases} 0; & x < a \\ \frac{x-a}{b-a}; & a \leq x \leq b \\ 1; & x \geq b \end{cases} \quad (4)_2$$

For the uniform distribution function as shown in Eq. (4), if we set $x = N$ and with the help of the technique described in this Section, we can obtain the following PDF and CDF in terms of given applied cycle N and damage level D .

$$f(D) = \begin{cases} N/[D^2(b-a)] & N/b \leq D \leq N/a \\ 0; & D > N/a \quad \text{or} \quad D < N/b \end{cases} \quad (5)_1$$

$$F(D) = \begin{cases} 0; & D < N/b \\ (Db - N)/[D(b-a)] & N/b \leq D \leq N/a \\ 1; & D > N/a \end{cases} \quad (5)_2$$

Intuitively, from Fig. 1 it seems like there is a close relationship between the failure distribution $f[N_f(D=1)]$ and $F[N_f(D=1)]$ and the damage distribution $f[D(N)]$ and $F[D(N)]$. Further examination shows that the two distributions do give the exactly the same reliability values when $D = 1$ for a given cycle N . For example, Eq. (3)₂ is exactly the same as Eq. (2)₂ and Eq. (5)₂ is exactly the same as Eq. (4)₂ when $D = 1$. The equivalency is held at least for the two distributions and under the linear damage assumption. Further mathematical analysis needs to be conducted to verify if the same reliability equivalency can be obtained for other distribution functions and nonlinear damage rules. The fact that the two reliability values are the same for a population is critical to the assumption used for the extended Binomial test method. More details will be provided in Section 3.2.

2.2. Discrete Failure and Damage Distributions for Samples

In practice it is difficult and often impossible to get the distribution of a population, in particular, for new product designs because of either budget constraints or timing or both. In the automotive industry, the number of samples used for product design and validation is usually limited. However, durability and reliability engineers have to find a way to effectively validate a product based on a limited sample size. The testing procedure with limited sample size is schematically illustrated in Fig. 2. The dashed distribution lines $f[N_f(D=1)]$ and $f[D(N)]$ indicate that the true population distributions are essentially unknown. The circles on the horizontal line $D = 1$ represent the cycles or time to failure of the samples. The solid circles stand for the samples below a specified cycle N , and the hollow circles stand for the samples above the specified cycle N . The diamonds on the vertical line at the specific cycle N represent the pass/fail status of the samples. The solid diamonds stand for pass, i.e. $D < 1$ and the hollow diamonds stand for fail, i.e. $D > 1$. How to validate a product from tests with limited sample size is a challenging and a critical issue. The

reliability criterion like Rxx alone is not good enough to be a design criterion, and the uncertainty caused by the sample size effect must be considered by any reasonable design criteria. The uncertainty can be better understood by conducting the following thought experiment: if a test with a certain sample size was repeated several times, the values of the distribution parameters obtained from fitting the test data would be different, sometimes, significantly different, especially for small sample sizes. Samples can seldom be taken to be representative of the whole population, especially for tests with very small sample sizes. Clearly, the uncertainty caused by the limited sample size or the associated confidence in tests must be addressed. Industry has developed some criteria over the years with the hope to capture the uncertainty effect and provide consistent rules in product validation. RxxCyy (e.g. R90C90 or R99C50) is such a criterion used in durability and reliability assessment for test data with limited sample sizes, for both the test-to-failure method and the Binomial test method. Again, R represents reliability, and the new parameter C stands for confidence.

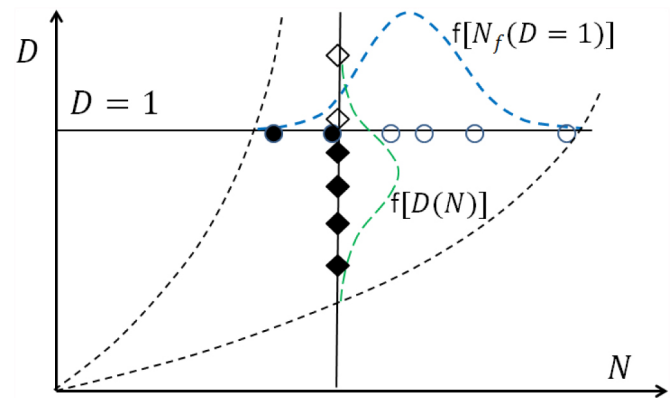


Fig. 2. Schematic of the discrete failure and damage distributions of the test data with limited sample sizes.

3. TESTING AND DATA ANALYSIS METHODS FOR PRODUCT VALIDATION

3.1. Test-to-Failure Method

Several approaches are available for estimating the RxxCyy value for failure test data. The two most commonly used methods are the tolerance limit approach and the probability plot approach.

Tolerance Limit Approach

With the tolerance limit approach [9, 10, 11, 12, 13], RxxCyy can be generally expressed as $Y_0 = \bar{Y} - Ks$ for test data so that a specific proportion R (100(1 - p)%) of the population will be above or below a prescribed confidence level C (100(1 - γ)%). Where $\bar{Y}$ and s are, respectively, the sample mean and sample standard deviation, which can be easily calculated. The only factor affecting the accuracy of the design curve is the value of K , which has closed-form approximate solutions under certain assumptions. Clearly, to obtain an accurate and reliable design curve, the accuracy of K value is critical. A specific K solution is [9, 11]: