SANDIA REPORT

SAND2012-7912
Unlimited Release
Printed September 2012

A New Approach to Quantification of Margins and Uncertainties for Physical Simulation Data

Justin T. Newcomer

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2012-7912
Unlimited Release
Printed September 2012

A New Approach to Quantification of Margins and Uncertainties for Physical Simulation Data

Justin T. Newcomer
Department of Human Factors and Statistics
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-MS0829

Abstract

This paper proposes the use of statistical tolerance interval methodology as an approach to quantification of margins and uncertainties (QMU) for physical simulation data. We review the standard k-factor methodologies and discuss potential limitations. The tolerance interval methodology is introduced and demonstrated with several examples. A new figure-of-merit is proposed and its properties are explored. These methodologies are intended for a performance characteristic that has shown the potential for low margin or margin that is changing with age. Hence, we require a well-understood dataset that has been through a comprehensive engineering analysis. This paper provides recommendations for an engineering analysis that will result in a dataset that is eligible for a rigorous analysis using these proposed methodologies. Finally, we present an overview of the probability of frequency approach commonly used in computational simulation QMU applications to highlight the similarities with this proposed methodology for physical simulation QMU applications.

ACKNOWLEDGMENTS

A large number of people have read and commented on early versions of this paper. The members of the QMU steering committee and the statisticians in organization 00431 were instrumental in completing this report. In particular, Brian Rutherford provided considerable insight in the development of the chapter on computational simulation and Tim Trucano and Stephanie Fitchett provided many thoughtful comments and feedback on several iterations of this paper.

CONTENTS

FIGURES

NOMENCLATURE

CA	component age
CD	compatibility definition
CDF	cumulative distribution function
CF	confidence factor
CV	critical value
FMEA	failure modes and effects analysis
k	k-factor
LPR	lower performance requirement
M	margin
QMU	quantification of margins and uncertainties
Q_r	r^{th} percentile
RMI	requirements modernization and integration
PC	performance characteristic
PDF	probability distribution function
PR	performance requirement
P_{req}	maximum allowable probability of failure
PS	product specification
TR	tolerance ratio
U	uncertainty
UPR	upper performance requirement

This page intentionally left blank

1. INTRODUCTION

Historically, Quantification of Margins and Uncertainties (QMU) for physical simulation test data has been centered on the calculation of a k-factor and/or a regression analysis of the k-factor against age. The k-factor, in general, is defined as margin divided by uncertainty. The most common interpretation of this definition, employed in a data rich environment, is where the margin is estimated by the difference between a defined requirement and the average response of the data and the uncertainty is estimated by the sample standard deviation of the data. These definitions are captured in the Annual Assessment RMI and are shown below.

- Metric: a characteristic of system or subsystem operation or performance that is, in principle, measurable.
- Threshold: a minimum or maximum allowable value of a given metric set by the responsible Laboratory.
- Best estimate: an assessed value of a given metric based on simulation, theory, or experiment.
- Margin: difference between the best estimate and the threshold for a given metric.
- Uncertainty: the range of potential values around a best estimate of a particular metric or threshold.
- K factor: figure-of-merit obtained from the ratio of component margins and their respective uncertainties based on standard deviations in data rich situations.

In a data rich environment most often the *metric*, as defined above, is interpreted to be the mean of the distribution of the system or subsystem operation or performance. The *threshold* is commonly defined to be a fixed requirement as defined in a product acceptance specification (PS) or a compatibility definition (CD). With this definition, one can then equate a computed k-factor directly to a probability that the system or subsystem will fail to meet its requirements (threshold) **provided the distribution of the system or subsystem operation or performance follows a Normal or Gaussian distribution**. Therefore, decisions are often driven by the size of the k-factor where cutoff or critical values are specified by percentiles of a Normal distribution. For example if $k > 2.576$ one could conclude that there is evidence that the probability that the system or subsystem will fail to meet its requirements is at most 0.005, since the 0.5^{th} (or the 99.5^{th}) percentile from a Normal distribution is 2.576 standard deviations from the mean. There are several issues with this definition of margin and hence the definition of the k-factor. Most notably, if there is a deviation from Normality the k-factor no longer relates directly to a probability of exceeding performance thresholds. The k-factor methodology can be a conceptual deviation from the general approach in a computational simulation framework and can fail to meet the key elements of applying a QMU methodology as defined in the Sandia Guidance Document on QMU [1] when the distribution of the system or subsystem operation or performance does not follow a Normal distribution.

The methodology proposed here is intended for a performance characteristic that has shown the potential for low margin or has shown that the margin is changing with age. This methodology shifts the focus of the analysis from the mean of the performance distribution to a meaningful percentile of the distribution. The percentile is chosen to correspond to a value that contains a certain acceptable proportion of the population units. This prompts the notion of margin to shift from the difference between the mean of a performance characteristic (PC) and its performance requirement (PR) to the difference between a meaningful percentile of the distribution of the performance characteristic and its performance requirement. It is also proposed to quantify uncertainty through the computation of a statistical confidence bound on the best estimate of the chosen percentile rather than by a sample standard deviation, which does not account for sampling variability. This is accomplished by computing a statistical tolerance interval. The tolerance interval differs from a confidence interval in that the confidence interval bounds a single-valued population parameter (the mean or the variance, for example) with some confidence, while the tolerance interval bounds the range of data values that includes a specific proportion of the population, with some level of confidence. A confidence interval's size is entirely due to sampling error, whereas a tolerance interval's size is due partly to sampling error and partly to actual variance in the population, and will approach the population's probability interval as sample size increases. The tolerance interval is also related to a prediction interval in that both put bounds on variation in future samples. The prediction interval only bounds a single future sample unlike a tolerance interval, which bounds the entire population (multiple future samples).

The remainder of this paper details a number of the limitations of the standard k-factor methodology and provides the details for an analysis using the newly proposed methodology. Examples are presented to reinforce these ideas. The proposed methodologies are intended for a thoroughly understood dataset with a performance characteristic that relates to component and/or system function and a well understood performance requirement. Much work is often needed to arrive at this state. This paper also outlines recommendations for an ***Engineering Data Analysis*** that will result in a dataset that is eligible for a rigorous QMU analysis. In addition, the proposed methodology is shown to be more consistent with the standard QMU methodologies for computational simulation applications.

2. K-FACTOR METHODOLOGY FOR PHYSICAL SIMULATION DATA

This section reviews the standard k-factor QMU methodology for physical simulation data. Here we solely focus on the statistical methodologies applied to a well-understood data set. There is a substantial amount of work involved in a preliminary graphical analysis, and in an engineering analysis of the data, that is omitted here. The engineering analysis is an essential step required to ensure the collected data sample includes measurements that may be used to infer performance in actual use. Some recommendations for an engineering analysis are presented in Section 5, and it is recommended that much greater emphasis be placed on doing this examination of the data well before the data analysis discussed here commences to ensure that both appropriate data and analysis techniques are used.

We assume that the measurement of the performance characteristics, as well as comparison to specified performance requirements, is sufficient to determine whether the component would have performed its intended function in actual use. Therefore, the performance requirement should be an accurate indication of a performance threshold. In the following discussion, we assume that there has been sufficient critical scrutiny (engineering analysis) performed to understand the performance characteristic and how it relates to component performance, the presented data set, and its representativeness of the larger population, and the performance requirements and their relationship to a performance threshold. For examples of the graphical analyses, engineering analysis, and the concepts discussed here refer to the QMU 101 course slides [2].

There are two main types of analyses performed that support QMU on physical simulation data. The first is a "*point-in-time*" analysis where the goal is to quantify the current amount of margin. This type of analysis is appropriate if data is collected at a single point in time, such as at product acceptance, or if data is collected over a range of component ages and there is no evidence of performance trending with age. If an age trend is detected then the second type of analysis is a "*k-factor regression analysis*" where the goal is to model the aging trend via statistical regression and identify an alarm age. The alarm age is defined to be the age at which the k-factor is no longer sufficiently large to assert that a desired proportion of components will be within the performance requirements with a given level of confidence. The following is an overview of each analysis.

2.1. Point in Time Analysis

A point in time analysis is performed if data is collected at a single point in time, such as at product acceptance, or if no trend is detected in the data. The goal of this analysis is to infer if there is a non-negligible probability that a component will fail to meet its performance requirements. Throughout this document, we refer to a ***Maximum Probability of Failure***, denoted by $\boldsymbol{P_{req}}$. This probability of failure specifically refers to a failure to meet performance requirements (margin failures), however we will simply use the terminology *maximum probability of failure* or P_{req}. This inference requires the quantification of the margin and uncertainty of the measured performance characteristic (PC) relative to an upper or lower performance requirement (UPR or LPR) at that point in time.

In the standard k-factor methodology, the margin is defined as the difference between the mean of the data and the performance requirement, $M = \mu - LPR$ or $M = UPR - \mu$. The uncertainty, U, in a data rich environment is often dominated by aleatory uncertainty (stochastic variability) and therefore is quantified by the standard deviation, σ, of the performance characteristic. In a data rich environment these two metrics, M and U, are can be characterized statistically by a probability density function (PDF) and its location relative to the performance requirement. Figure 2.1 below depicts these metrics relative to a Normal distribution function. These metrics are defined in the figures below, and the equations that follow, relative to the lower performance requirement (LPR) for reference. For an upper bound, the metrics could be appropriately adjusted.

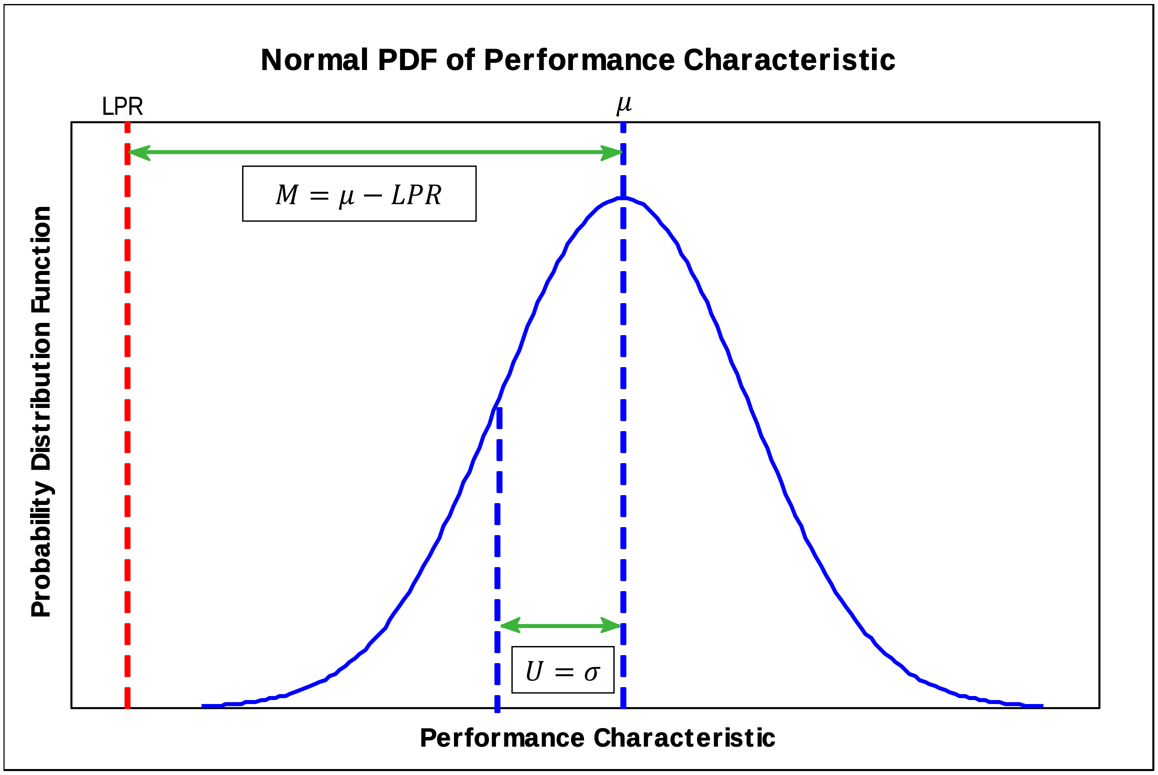

Figure 2.1. Graphical Depiction of M and U Relative to a Normally Distributed PC.

To estimate these quantities a representative random set of data is collected. Figure 2.2 and Figure 2.3 below show an example of a measured performance characteristic for a sample of units (Example Dataset 1) relative to a lower performance requirement (LPR).

The scatter plot in Figure 2.2 is a useful tool to assess if an age trend is present. The green dashed line represents an estimated regression fit on the data. Although the data appears to trend slightly upward, the slope of the regression fit is not significantly different from zero (p-value[1] = 0.575). Since there is not a statistically significant trend, a point-in-time analysis is appropriate. Another useful plot, when no trend is present, is a histogram of the observed performance characteristic (Figure 2.3). This plot provides a snapshot of the distribution of the performance characteristic and is useful in assessing (at least visually) if Normality is a viable assumption.

[1] In hypothesis testing, a p-value represents the probability of a test statistic obtaining a value at least as extreme as the observed value derived from a sample. When a p-value is lower than the significance level of a test, the test is considered to be significant and the null hypothesis is rejected.

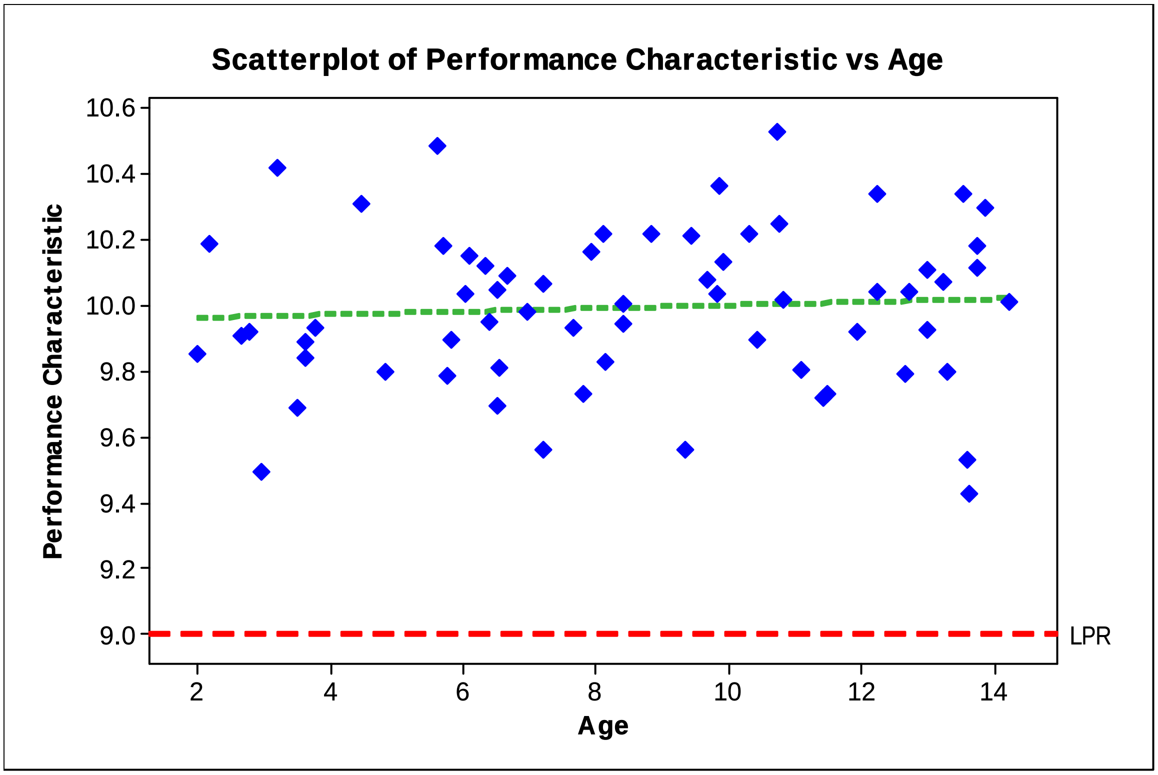

Figure 2.2. Scatter Plot of Example Dataset 1 Showing No Aging.

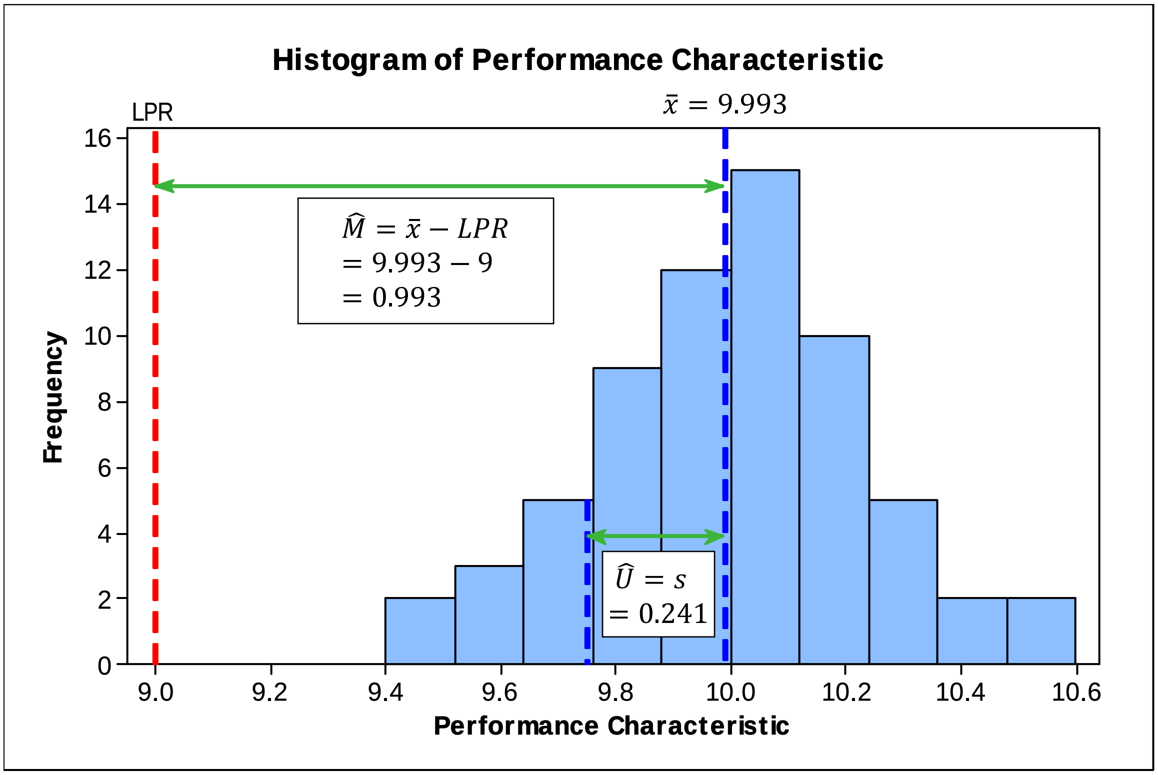

Figure 2.3. Histogram of Example Dataset 1 Showing the K-factor QMU Metrics.

To assess overall margin relative to the performance requirement and uncertainty the figure of merit for a point-in-time analysis is the k-factor. The k-factor is defined as the ratio of the margin to the uncertainty,

$$k = \frac{M}{U} = \frac{\mu - LPR}{\sigma}.$$

Given a sample of data the mean and standard deviation can be estimated by the sample mean, $\bar{x}$, and sample standard deviation, s. For a sample of size n these are defined as,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Hence, the best estimate of the k-factor is given by,

$$\hat{k} = \frac{\widehat{M}}{\widehat{U}} = \frac{\bar{x} - LPR}{s}.$$

This estimate gives a measure of the estimated margin relative to the performance requirement in terms of the estimated standard deviation. For the data shown in Figure 2.3 there were 65 observations yielding an estimated mean and standard deviation of $\bar{x} = 9.993$ and $s = 0.241$. Therefore, the estimated k-factor for this data, relative to a lower performance requirement of 9 is

$$\hat{k} = \frac{\bar{x} - LPR}{s} = \frac{9.993 - 9}{0.241} = 4.12.$$

The k-factor itself does not directly estimate the probability of a performance characteristic failing to meet the performance requirements. However, for a Normal or Gaussian population the k-factor can be used to estimate the probability of failing to meet the lower performance requirement by,

$$P = Prob(PC < PR) = \Phi\left(\frac{LPR - \bar{x}}{s}\right) = 1 - \Phi(\hat{k}),$$

where $\Phi(\cdot)$ is the cumulative probability function for a standard Normal distribution ($\mu = 0$, $\sigma = 1$). Suppose there is a requirement that this probability P must be less than a specified maximum probability of failure, P_{req}. This gives us the following relationship,

$$P = 1 - \Phi(\hat{k}) < P_{req} \Leftrightarrow \hat{k} > \Phi^{-1}(1 - P_{req}).$$

For example, suppose we require $P < P_{req} = 0.005$, then the critical value (CV) for the k-factor is given by $CV = \Phi^{-1}(0.995) = 2.576$. Therefore, if the estimated k-factor is greater than 2.576 we can conclude that there is evidence that the probability of obtaining a performance characteristic less than the performance requirement is at most 0.005. A similar calculation can be performed for an upper performance requirement. For Example Dataset 1, we estimated the k-factor to be $\hat{k} = 4.12$. Since 4.12 > 2.576 we conclude there is evidence that the probability of obtaining a performance characteristic less than the lower performance requirement of 9 is at most 0.005. Similarly, the probability that an observed performance characteristic will be greater than the requirement of 9 is estimated to be at least 0.995.

This evidence however is based on a single sample of data from a larger population. Since the goal of QMU is generally to estimate the probability of failing to meet the performance requirements *with a specified level of statistical confidence*, to account for the uncertainty from the sample itself one should calculate a statistical lower confidence bound on the k-factor. Then, if the $\gamma \cdot 100\%$ lower confidence bound on the k-factor, denoted by $\hat{k}_\gamma$, is greater than CV we could conclude that we are $\gamma \cdot 100\%$ confident that the probability of obtaining a performance characteristic less than the performance requirement is at most P_{req}. Several methods for obtaining confidence bounds on the k-factor are available and are presented in the QMU 102 course material [3]. **These methods in general assume that the data are a random sample from a Normal population**. Issues arising due to a deviation from this assumption are discussed in Section 3.1. For Example Dataset 1, for which Normality is a reasonable assumption, an estimated 95% lower confidence bound on the k-factor is $\hat{k}_{0.95} = 3.48$. Since 3.48 > 2.576 we can assert that we have 95% confidence that the probability of obtaining a performance characteristic less than the lower performance requirement of 9 is at most 0.005.

2.2. K-Factor Regression Analysis

A k-factor regression analysis is performed if the observed data appears to be trending with age. Again, the identification of the age trend, along with a determination that the trend is due to an aging effect and not some other known or unknown factor, should be done during the engineering analysis of the dataset. Note that a performance characteristic may also trend with respect to other factors, such as temperature, voltage, etc. The methodologies discussed here could be applied to these factors as well, however we will only focus on an aging trend for example purposes.

This section describes the statistical analysis needed to estimate an alarm age. The ***Alarm Age*** is defined as the component age at which we estimate certain percentage of the population is no longer contained by the performance requirement, with a given level of confidence. Measured values of a performance characteristic may trend up or down, or the range of values may grow larger or smaller as components age. We are most concerned if the measured values trend toward a performance requirement or if the range of measured values expands toward the limit. Figure 2.4 shows a notional plot of the mean performance characteristic versus component age (CA). Here the mean, $\mu(CA)$, and the margin, $M(CA) = |\mu(CA) - PR|$, are functions of the component age. The uncertainty is quantified by the standard deviation around the fitted regression line, σ_R, which may also be a function of the component age. The notional data in Figure 2.4 depicts a downward trend and increasing measurement error bars with increasing component age. The LPR is shown by the red dashed line. The plot thus demonstrates an example of decreasing performance with increasing variability, suggesting possible failure at some point in the future. Note, the case of increasing variability with age is shown here for example purposes only. The regression analyses described throughout this paper assume a constant variance. For data that exhibits increasing variability with age, a statistician should be consulted to determine the appropriate methodology to apply.

As discussed in Section 2.1, a scatter plot is a useful tool to assess if an aging trend is present. Figure 2.5 shows a scatter plot of Example Dataset 2 with an aging trend. It is possible that an observed visual trend will not be statistically significant. That is, the slope of the fitted regression line may not be significantly different from zero or the variability in the data may be too large to conclude with confidence that the trend is present. The green dashed line in Figure 2.5 is the estimated regression line for Example Dataset 2. For simplicity we will only explore a linear regression model with the component age (CA) as a single independent variable. The regression model is given by,

$$PC = \beta_0 + \beta_1 \cdot CA + \varepsilon,$$

where PC is the performance characteristic, β_0 and $\beta_1$are the model parameters to be estimated, and ε is a random error assumed to follow a Normal distribution with mean zero and standard deviation σ_R.

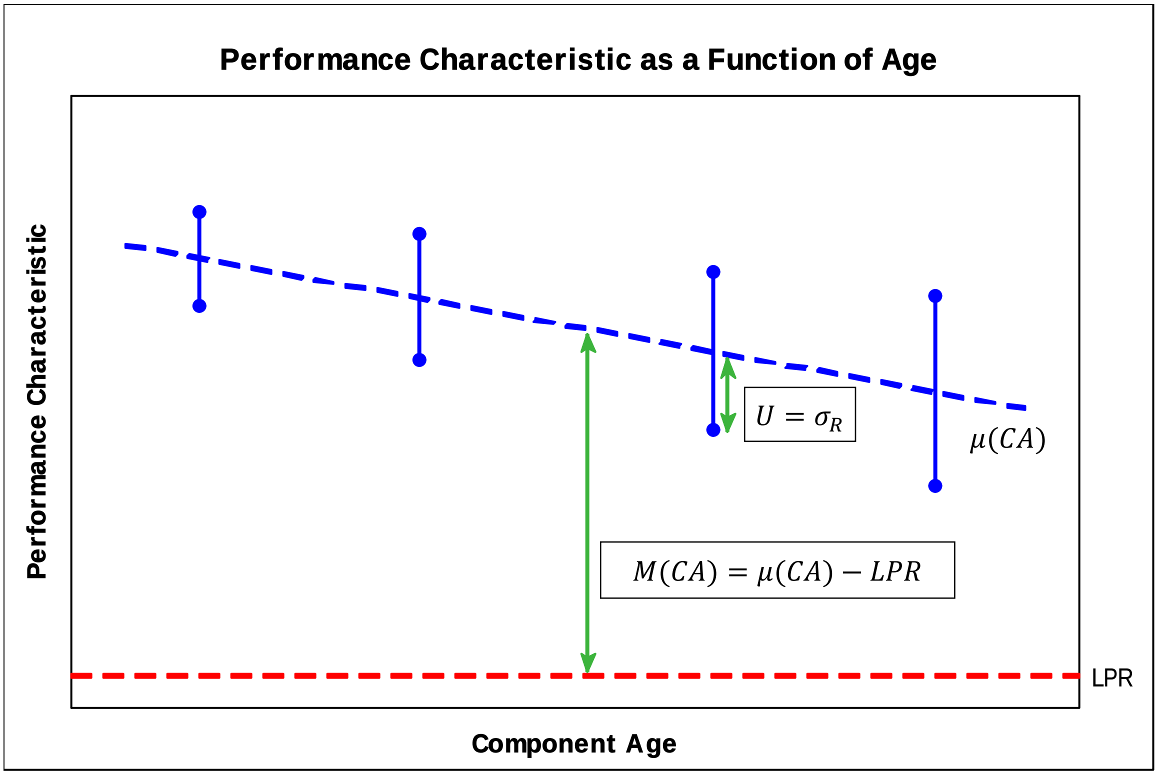

Figure 2.4. Graphical Depiction of Margin and Uncertainty with a Linear Aging Trend.

For Example Dataset 2, shown in Figure 2.5, the model parameters are estimated to be $\hat{\beta}_0 = 10.493$ and $\hat{\beta}_1 = -0.031$ which gives the estimated regression line for the mean performance characteristic at age CA, shown by the dashed green line, to be $\widehat{PC} = 10.493 - 0.031 \cdot CA$. A statistical test can be performed to assess if the estimated slope $\hat{\beta}_1$ is significantly different from zero (i.e. the aging trend is statistically significant). For Example Dataset 2, the test yields a p-value of 0.001, which indicates that the slope is significantly different from zero. Further, we can obtain an estimate of σ_R which is the standard deviation around the regression line or the standard deviation of the residuals[2] of the regression fit. For a dataset with an aging trend and constant variance, this is the measure of uncertainty. Example Dataset 2 yields an estimate of $\hat{\sigma}_R = 0.246$. This initial regression analysis is performed to assess the significance of the aging trend and to obtain an estimate of uncertainty given the assumed linear regression model.

[2] The residuals are observed errors of the fitted regression model, $PC_i - \widehat{PC}_i$, which are empirical estimates of the regression model errors, ε_i, which are not observed.

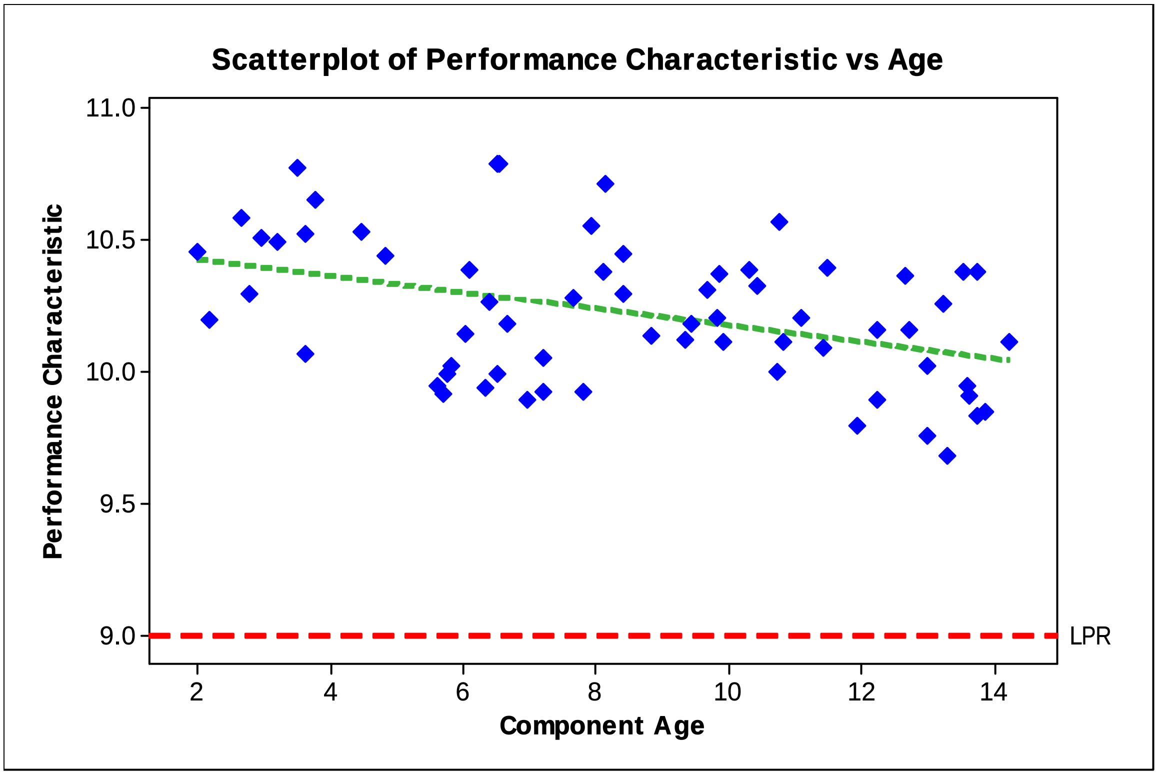

Figure 2.5. Plot of Example Dataset 2 Showing and Aging Trend and Regression Fit.

The second step is to perform a regression on the k-factor. First we must transform the dataset to a standardized k-factor scale. This is accomplished in a similar fashion to the computation of the estimate of the k-factor in a point-in-time analysis, however here the calculation is performed on each individual data point. Therefore, we call these standardized quantities "*individual k-factors*" and they are defined for a lower requirement as,

$$k_i = \frac{PC_i - LPR}{\hat{\sigma}_R},$$

where the subscript i denotes the ith observation or unit. These factors, by their very nature, are population related; that is, the standardization by the population statistic $\hat{\sigma}_R$ means that these quantities are summaries of population characteristics. These standardized quantities can be referenced to probability-of-compliance levels through the assumption that the distribution of the responses is approximately Normal around the age-regressed mean.

Recall, we have defined the alarm age to be the component age at which we estimate certain percentage of the population is no longer contained by the performance requirement, with a given level of confidence. This is accomplished by estimating a confidence interval around the mean regression line with respect to the age-trended individual k-factors. Again, a critical value (CV) must be chosen to correspond to a maximum probability of failing to meet the performance requirement. For this example we again assume the requirement $P < P_{req} = 0.005$ and the critical value for the k-factor is given by $CV = \Phi^{-1}(0.995) = 2.576$. Figure 2.6 shows the regression analysis of the individual k-factors versus age with an estimated 90% confidence interval (red dashed lines) around the mean k-factor regression line (green dashed line).

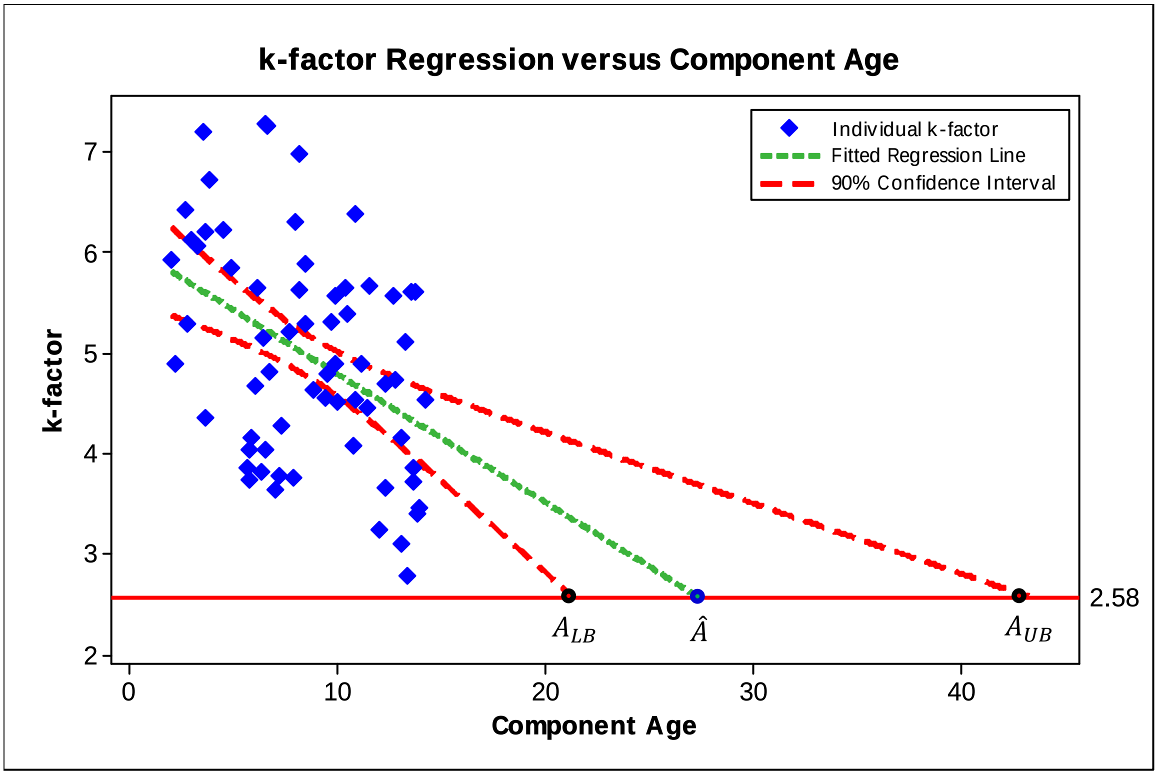

Figure 2.6. K-factor Regression of Example Dataset 2.

The mean line in Figure 2.6 (shown in green) is interpreted as the expected value of the k-factor as a function of age. The predicted alarm age, $\hat{A}$, is calculated by determining the age at which the mean trend line crosses the critical value line (red solid line). The ages at which the confidence bounds cross the reference line provide confidence bounds for the predicted alarm age. These confidence bounds are denoted by the interval (A_{LB}, A_{UB}). For Example Dataset 2, $\hat{A} = 27.3$ years, and the confidence bounds are (A_{LB}, A_{UB}) = (21.3 years, 43.2 years). The lower bound of the 90% confidence interval, A_{LB}, is a one sided 95% lower confidence bound for the predicted alarm age. This is the age at which it is predicted that $\hat{k}_{0.95} = 2.576$. Therefore, for components with ages greater than $A_{LB} = 21.3$ we cannot assert that we have 95% confidence that the probability of obtaining a performance characteristic greater than the lower performance requirement of 9 is at least 0.995. Figure 2.6 can be difficult to explain to customers. Further, the confidence bounds on the k-factor regression line do not translate back to the engineering unit scale easily. Therefore, it has been common practice to show an additional graphic of the original regression line with ***Prediction Intervals***. This is shown in Figure 2.7 below. In cases with large sample sizes the point at which the lower prediction interval crosses the lower performance requirement can be similar to the alarm age estimate from the k-factor regression line. These intervals however, do not relate to the confidence intervals on the k-factor regression line (as can be seen by comparing Figure 2.6 and Figure 2.7) and can add further confusion. This will be discussed further in Section 3.2.

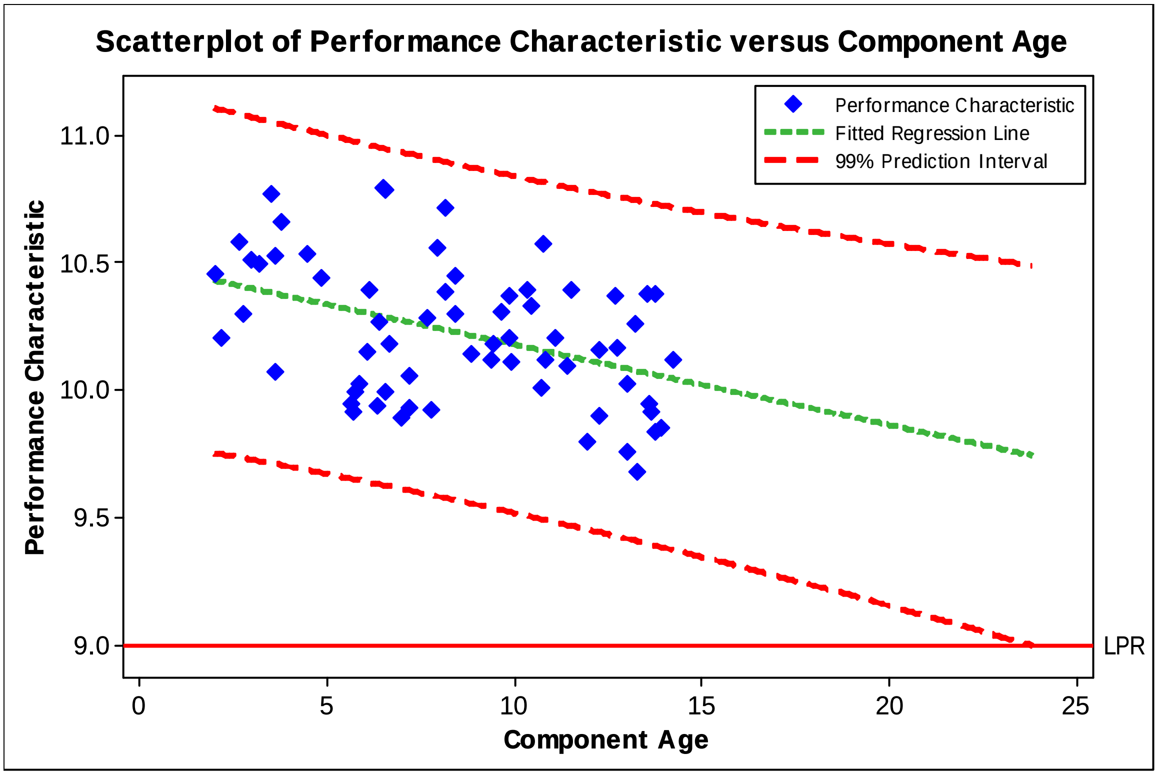

Figure 2.7. Plot of PC vs. CA with Regression Fit and 90% Prediction Interval

This methodology, and the new methodology proposed in Section 4, requires a projection of the regression line beyond the range of observed component ages. This should be done cautiously and ideally only when the aging trend is understood well enough, that we believe the linear model to be justified beyond the timeframe for which the data are available. It should also be noted that the computation of the confidence bounds in the age trend methodology are confidence bounds on the mean k-factor regression line and do not account for the uncertainty in the estimation of the standard deviation around the regression line (which feeds into the estimation of the individual k-factors). This issue will be discussed in more detail in Section 3.4. The new methodology proposed in Section 4 gives an approach that accounts for the uncertainty in the estimation of both the mean and standard deviation.

3. LIMITATIONS OF THE K-FACTOR METHODOLOGY

This section discusses several issues and concerns with the standard k-factor QMU methodology for physical simulation data. Most issues apply to both the point-in-time analysis and the k-factor regression analyses; however we will only illustrate the issues using the point-in-time analysis.

3.1. Non-Normal Data

One of the main difficulties with the k-factor methodology arises when the assumption of Normality is violated. When this situation arises, the k-factor is no longer a good indicator of the probability of failing to meet performance requirements (i.e. the demonstrated performance). This is illustrated in Figure 3.1 below, which shows a Normal distribution function compared to a Weibull distribution function with exactly the same mean and standard deviation. Therefore, the estimated k-factors would be the same for these two distributions. It is clear however that the tails of these distributions at the lower requirement are quite different. If we incorrectly use the Normal distribution, we would underestimate the probability of failure (area under the curve to the left of the lower performance requirement). Moreover, for inferences with respect to the upper tails the Normal distribution may overestimate a probability of failure for an upper requirement.

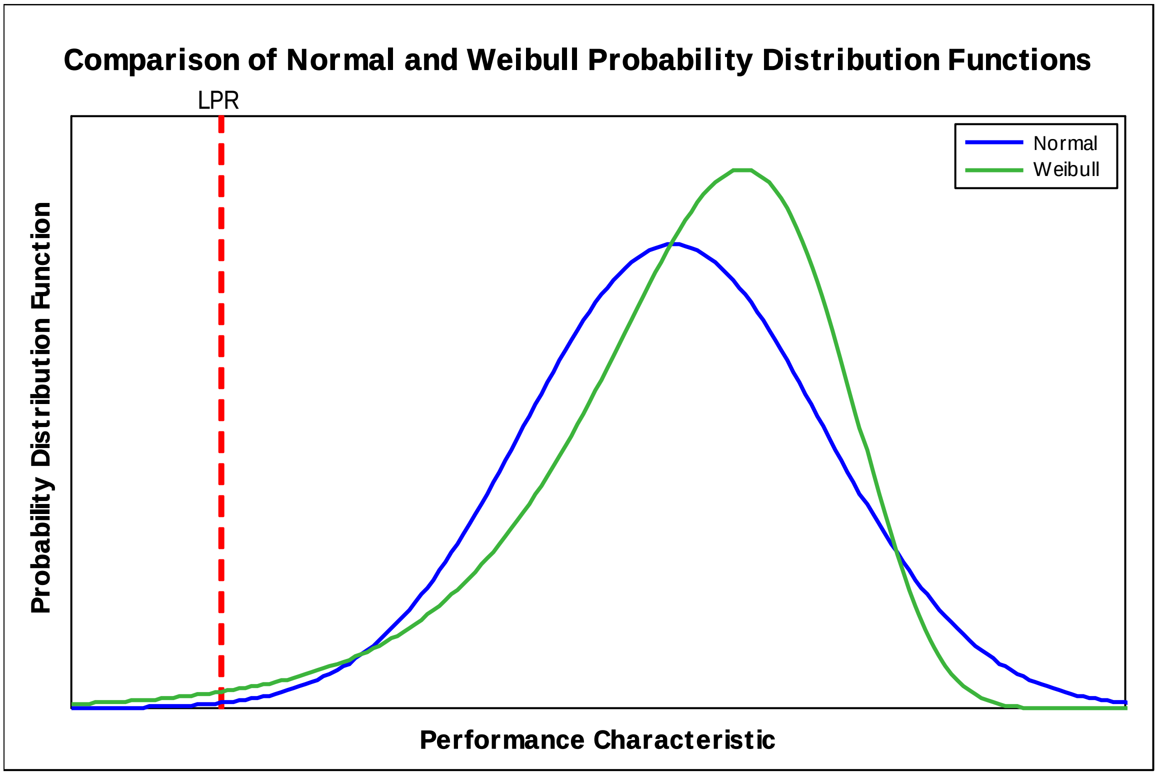

Figure 3.1. Normal and Weibull Distributions with the Same Mean and Standard Deviation.

For example, consider the data shown in Figure 3.2 below (Example Dataset 3). Clearly, the assumption of Normality is not valid and the data appear to be skewed toward the lower performance requirement.

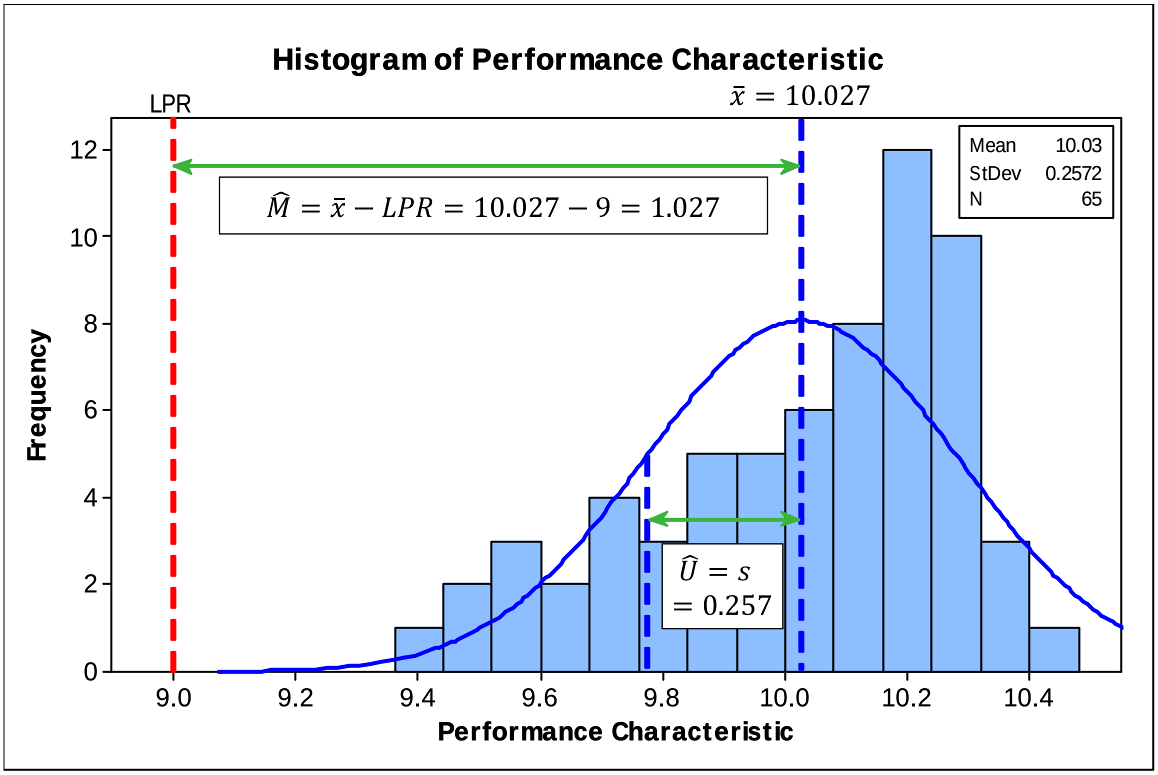

Figure 3.2. Histogram of Example Dataset 3 Showing Non-Normality.

Example Dataset 3 consists of 65 observations of the performance characteristic yielding an estimated mean and standard deviation of $\bar{x} = 10.027$ and $s = 0.257$. Therefore, the estimated k-factor for these data, relative to a lower performance requirement of 9 is estimated to be,

$$\hat{k} = \frac{\bar{x} - LPR}{s} = \frac{10.027 - 9}{0.257} = 3.99.$$

Further, a 95% lower confidence bound on the k-factor is estimated to be $\hat{k}_{0.95} = 3.38$. Since $3.38 > 2.576$ we could assert that we have 95% confidence that the probability of obtaining a performance characteristic less than the lower performance requirement of 9 is at most 0.005. This assertion however requires that we assume that the data follow a Normal distribution. Figure 3.2 clearly shows that the Normal distribution fit is not appropriate and the lower tail of the fitted distribution is smaller than the observed data. We will show using the new methodology that the k-factor and confidence bound estimated on this dataset are overly optimistic and ultimately lead to incorrect conclusions.

To verify or refute an assumption of Normality, one could produce a Normal probability plot for the data in Figure 3.2. Figure 3.3 below shows several probability plots that reinforce that Normality is not a valid assumption. A common approach that analysts take when the Normality assumption is violated is to transform the data to Normality and then analyze the transformed data. Figure 3.3 indicates that the Weibull or Smallest Extreme Value distributions or a Johnson Transformation [4] might be appropriate choices for this dataset. Since the Weibull and Extreme Value distributions do not have a one-to-one relationship with the Normal distribution, the Johnson Transformation is often used for its convenience.

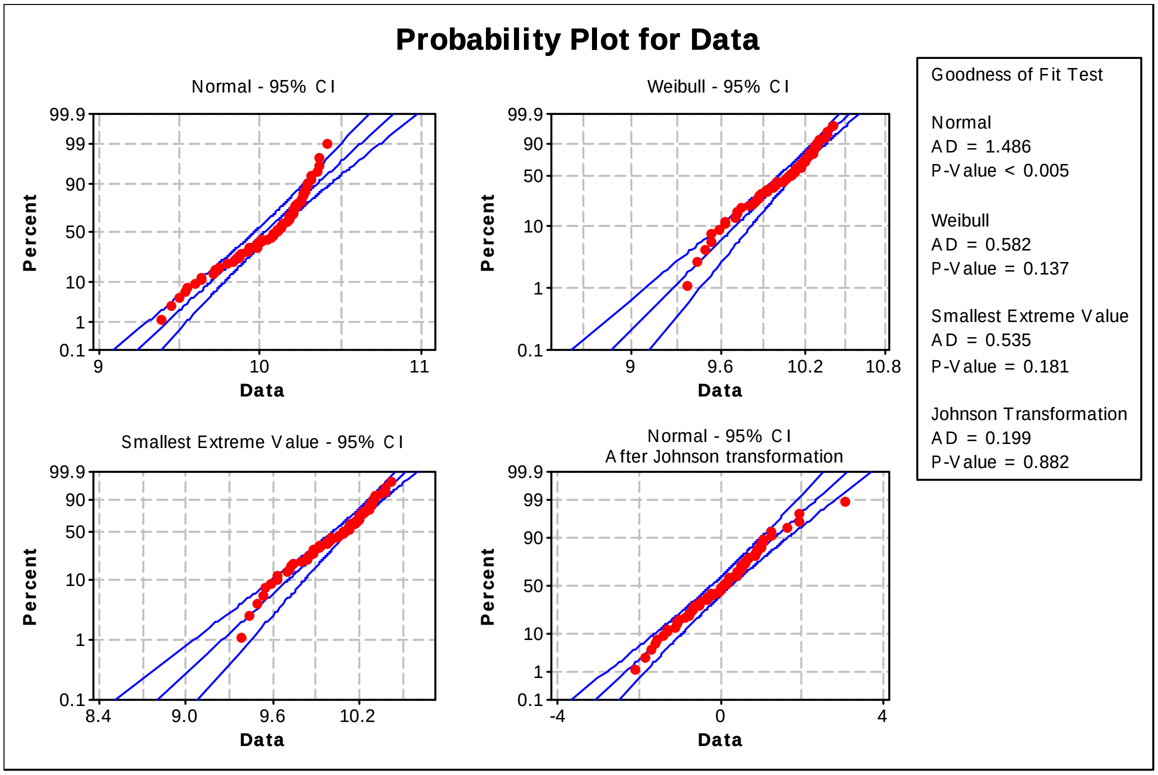

Figure 3.3. Probability Plots for the Non-Normality Example Dataset 3.

The best fit Johnson Transformation for this dataset is given by,

$$X_T = -1.098 + 0.990 \cdot \ln\left(\frac{X - 8.983}{10.435 - X}\right).$$

Therefore, a k-factor and confidence bound can be computed on the transformed data relative to a transformed lower performance requirement of

$$LPR_T = -1.098 + 0.990 \cdot \ln\left(\frac{LPR - 8.983}{10.435 - LPR}\right) = -5.47.$$

Figure 3.4 shows the transformed data and transformed lower performance requirement. From the transformed data one can get a revised estimate of the k-factor and confidence bounds. For this transformed data set, we obtain an estimated mean and standard deviation of $\bar{x}_T = 0.024$ and $s_T = 1.003$. Therefore, the estimated k-factor for this data, relative to a transformed lower performance requirement of -5.47 is estimated to be,

$$\hat{k}_T = \frac{\bar{x}_T - LPR_T}{s_T} = \frac{0.024 - (-5.47)}{1.003} = 5.48.$$

Further, a 95% lower confidence bound on the transformed k-factor is estimated to be $\hat{k}_{0.95,T} = 4.65$. Since 4.65 > 2.576 we could assert that we have 95% confidence that the probability of obtaining a **transformed** performance characteristic less than the **transformed** lower performance requirement of -5.47 is at most 0.005. Although in some cases this may give us a better estimate of the demonstrated performance, it does not directly translate to the probability of an observed performance characteristic on the original scale being greater than the original scale lower performance requirement. That is, $P_T < P_{req} \not\Leftrightarrow P < P_{req}$, where P_T is the probability of obtaining a **transformed** performance characteristic less than the **transformed** lower performance requirement. This relationship may hold under some simple (e.g. linear) transformations, but in general, the best fitting transformations are of a more complicated form (as is the case shown here). We will show in Section 4.3.2 that this is an overly optimistic and misleading result.

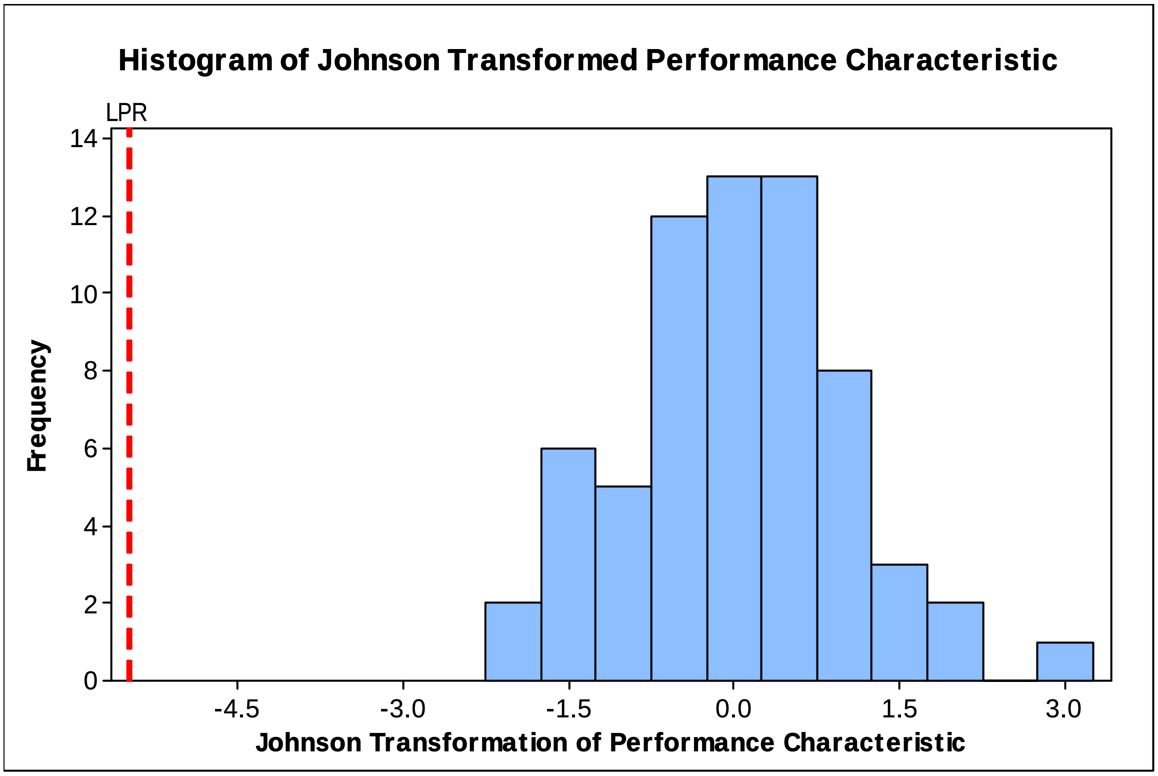

Figure 3.4. Histogram of Example Dataset 3 after applying a Johnson Transformation.

If the violation of the Normality is ignored and the data is analyzed on the original scale, the estimate of the k-factor no longer gives an accurate measure of demonstrated performance and can be misleading and result in erroneous conclusions. The new methodologies, presented in Section 4, provide a way to estimate the demonstrated performance directly and present the results and conclusions all on the original engineering unit scale.

3.2. Interpretability

Another limitation of the k-factor methodology, especially when the assumption of Normality is violated, is that the interpretability becomes extremely compromised. The k-factor estimated from transformed data does not translate back to the original scale and therefore is difficult to explain. Further, by transforming the data and presenting results on a transformed scale, useful information about the shape of the distribution can be lost. Presenting graphics, such as the one in Figure 3.4, to customers and decision makers can be misleading and difficult to explain properly. This leads to a decreased understanding of the conclusions, their impact to performance, and the assumptions that are required to make those conclusions.

Another misleading graphic is the k-factor regression plot shown in Figure 2.6. There are several metrics included on this regression plot that are difficult to understand and explain to customers and decision makers. First, the individual k-factors are population quantities due to the nature of their computation; however, they take the appearance of individual observations. Because of this, the k-factor regression is always decreasing with component age, even if the performance characteristic distribution is increasing (toward an upper requirement). This can be conceptually confusing. Further, the critical value shown on this plot is merely a percentile of a standard Normal distribution and has no relationship to the performance requirement. The performance requirement rather is captured in the individual k-factors, which further complicates the explanation of this graphic. Finally, the confidence bounds on the k-factor regression line, which are used to estimate uncertainty in the estimate of the alarm age, do not translate back to the engineering unit scale easily. Therefore, a second graphic, shown in Figure 2.7, is often presented to customers and decision makers to demonstrate these results on the engineering unit scale. The point at which the prediction limit crosses the lower performance requirement is sometimes used as an estimate for the alarm age, and this value can be quite similar to the k-factor regression estimate of the alarm age in cases with large sample sizes. **However, the interpretation of this estimate from a prediction interval is for a single future observation and not for the population**. Further, the prediction interval technique does not provide a way to obtain confidence bounds on the alarm age. The similarities and differences in these two intervals is difficult to explain and can be misleading. Further, as mentioned in Section 1, the prediction interval only captures information about a single future observation. In Section 4, we introduce a new methodology based on the computation of a statistical tolerance interval. A tolerance interval covers a specified proportion of a population with a certain confidence level, making the tolerance interval more appropriate if a single interval is intended to bound a population of units or multiple future samples. This interval is computed and presented on the engineering unit scale making only a single graphic necessary to present the results and conclusions.

3.3. Critical Value and Decision Rule

Another limitation of the standard k-factor methodology is that although the k-factor itself is a standardized quantity it does not take into account the desired performance that is to be demonstrated (i.e. the probability of failing to meet performance requirements should be at most P_{req}). This rather is accomplished by specifying a suitable critical value corresponding to P_{req}. This is easily demonstrated using Example Dataset 1. Recall for this dataset we had $\hat{k} = 4.12$ and $\hat{k}_{0.95} = 3.48$. Therefore, if we would like to demonstrate that the maximum probability of observing a performance characteristic less than the lower performance requirement of 9 is at most $P_{req} = 0.005$, then we would use a critical value of $\Phi^{-1}\left(1 - P_{req}\right) = \Phi^{-1}(0.995) = 2.576$. Since the assumption of Normality is sufficient for this dataset and $\hat{k}_{0.95} = 3.48 > 2.576$ we could conclude that we have 95% confidence that the probability of obtaining a performance characteristic less than the lower performance requirement of 9 is at most 0.005. If we assume however that this performance characteristic must have a more stringent requirement on the maximum probability of failure then our critical value will change. This can occur on components that have several potential failure mechanisms that are monitored by several performance characteristics. Then, to demonstrate an overall probability of failure each individual performance characteristic must demonstrate meeting a more strict maximum probability of failure. Suppose the requirement for this example is that the maximum probability of observing a performance characteristic less than the lower performance requirement of 9 is at most $P_{req} = 0.0001$. The critical value for this new requirement becomes $CV = \Phi^{-1}(0.9999) = 3.719$. Given the estimated 95% lower confidence bound on the k-factor of $\hat{k}_{0.95} = 3.48$ we cannot claim that we are meeting the requirement with 95% confidence.

It is common for different components to have different requirements and it may even be the case that different performance characteristics from the same component could have different requirements.

Therefore, it can become quite confusing when k-factors from one performance characteristic are compared against one critical value while another is compared against a different critical value. This induces the common question about the k-factor of "what is large enough" and causes the answer to be overly confusing. Further, this is conceptually very different than the framework in computational simulation applications where $M/U > 1$ indicates that the requirements are being met. In the methodology proposed in Section 4 the required probability of failure is incorporated into the calculation of a new margin divided by uncertainty figure-of-merit. It will be shown that this allows us to use a single critical value for all performance characteristics regardless of their requirements.

3.4. Quantification of Sampling Uncertainty

In the standard k-factor methodology, the uncertainty is quantified by the sample standard deviation. This provides an **estimate** of the stochastic variability in the measurements of the performance characteristic. This aleatory uncertainty accounts for variability arising from a number of effects (unit-to-unit differences, lot-to-lot differences, lack of measurement precision, etc.), none of which should alone account for the majority of the total stochastic variability. The estimate of the k-factor is a function of the estimated margin and uncertainty. It is, however, just an estimate based on a single sample of data points. If a different sample of data were chosen the estimates of margin, uncertainty, and hence $\hat{k}$ would be different. The uncertainty associated with the random variations in the estimates of population parameters is referred to as ***Sampling Uncertainty***. An estimate of the k-factor, based upon the mean and standard deviation from a single sample of data clearly does not quantify this type of uncertainty.

A range of possible values of a population parameter may be computed from the sample with a defined likelihood (a statistical confidence level) of bounding the true population parameter. This range is called a (statistical) confidence interval, and is usually labeled with the confidence level. Confidence intervals for high levels of confidence are wider than confidence intervals for lower levels of confidence. Therefore, as discussed in Section 2.1, a lower confidence bound on the k-factor should be computed to account for this additional uncertainty. The computation of this confidence bound however is not trivial and in most cases requires an approximation or simulation based technique to obtain the bound. Further, these approaches in most cases are only applicable to Normal data. A commonly used simplification to obtain a confidence bound is to assume the standard deviation is a known constant and obtain a confidence bound that accounts for uncertainty in the mean only. This is the case in the k-factor age trend regression analysis. The individual k-factors are regressed versus age and a confidence interval is obtained on the **mean** k-factor. This requires the assumption that the standard deviation is known and hence the confidence interval only accounts for uncertainty in the estimation of the mean of the data.

The approach described in Section 4 defines uncertainty to account for sampling uncertainty. We will also provide an example of how this approach can be easily applied to a distribution other than Normal. Finally, the choice of the confidence level will be incorporated into the estimate of the new figure-of-merit without having to redefine a new critical value.

4. A NEW APPROACH TO QMU FOR PHYSICAL SIMULATION DATA

The methodology proposed here is intended for a performance characteristic that has shown the potential for low margin, a deviation from the assumption of Normality, or that the margin is changing with age. This methodology shifts the focus of the analysis from the mean of the performance distribution to a meaningful percentile of the distribution. The percentile is chosen to correspond to a value that contains a certain acceptable proportion of the population units. This prompts the notion of margin to shift from the difference between the mean of a performance characteristic (PC) and its performance requirement (PR) to the difference between a meaningful percentile of the distribution of the performance characteristic and its performance requirement. It is also proposed to quantify uncertainty through the computation of a statistical confidence bound on the best estimate of the chosen percentile rather than by a sample standard deviation. Further, this approach will bridge a current conceptual gap between QMU for physical simulation and QMU for computational simulation applications.

Again, the methodologies presented here are intended for a well-understood dataset that has been through a comprehensive engineering analysis. This newly proposed methodology incorporates both the required maximum allowable probability of failure and the confidence level into the computation of the new margin divided by uncertainty figure-of-merit. Therefore, these essential pieces of information must be discussed in the engineering analysis and documented **prior to** performing these more rigorous statistical analyses. We assume that a dataset that has been through a rigorous engineering analysis has the following properties.

1. A subset of the performance characteristics have been identified to have an impact on performance or are meaningful for successful component or system function.

2. The performance requirements for each performance characteristic to be analyzed are known and have a meaningful engineering justification.

3. The pedigree and representativeness of the units tested are understood and relate directly to a known population of interest. In addition, the testing procedures are known to be an accurate representation of use conditions and component function and should be well documented.

4. The existence and impact of all potential factors (environmental conditions, lot-to-lot differences, launch profiles, tester differences, etc.) are known to a degree that allows one to account for them properly in the analysis.

5. The sample size and quality of the dataset are sufficient for further analysis. A dataset with a small sample size, unknown factors, or large measurement uncertainty may not be appropriate for these methodologies. A statistician can provide guidance to decide if a dataset meets this criterion.

Further, upon completion of the engineering analysis of the data, an *Engineering Review* should be performed by an interdisciplinary team (Systems and Components Engineering, Test Engineers, Statisticians and Analysts, etc.) to ensure that the conclusions resulting from the engineering analysis are acceptable and the dataset is sufficient for further analysis.

4.1. Overview and Definitions

The methodologies proposed in this paper are intended to answer the following questions:

1. Are we $\boldsymbol{YY}$% certain that **at-least** $\boldsymbol{XX}$% of the unit population will yield a response **greater than** the threshold $\boldsymbol{T}$?

2. Are we $\boldsymbol{YY}$% certain that **at-least** $\boldsymbol{XX}$% of the unit population will yield a response **greater than** the threshold $\boldsymbol{T}$ after $\boldsymbol{Z}$ years of life?

3. At which age will we no longer be $\boldsymbol{YY}$% certain that **at-least** $\boldsymbol{XX}$% of the unit population will yield a response **greater than** the threshold $\boldsymbol{T}$?

The values of $\boldsymbol{XX}$, $\boldsymbol{YY}$, $\boldsymbol{Z}$, and $\boldsymbol{T}$ and the comparisons '**at-least**' versus '**at-most**' and '**greater than**' versus '**less than**' are all parameters of the requirement. This requirement reflects a common question addressed through QMU. Note that, in some applications the appropriate metric (XX) could be defined as the median or as the mean (as in the k-factor methodology) of the performance characteristic distribution.

A measure of demonstrated performance is a ***Percentile*** of the distribution of the performance characteristic corresponding to the ***Maximum Allowable Probability of Failure,*** $\boldsymbol{P_{req}}$. As discussed in Section 2.1, this probability of failure specifically refers to a failure to meet performance requirements (margin failures), however we will simply use the terminology *maximum probability of failure* or P_{req}. A percentile of a distribution is defined as the value of a variable (here the performance characteristic) below which a certain percent of the values for that variable will fall. We denote a percentile by Q_r, where r represents the probability that a performance characteristic value will fall below Q_r. That is,

$$Prob(PC < Q_r) = r.$$

For inferences with respect to a lower requirement, we are generally concerned with the lower tail and percentile of the performance characteristic distribution and for inferences with respect to an upper requirement, we are concerned with the upper tail and percentile.

We define the ***Content*** of a distribution to be the proportion of units that are expected to be within the performance requirements (proportion greater than a lower requirement or proportion less than an upper requirement) and denote this quantity by lower case p. Recall question 1 above; are we YY% certain that at-least XX% of the unit population will yield a response less than the threshold T? Here XX% corresponds to the desired content, $XX\% = p \cdot 100\%$. Further, the content is one minus the maximum allowable probability of failure, $p = 1 - P_{req}$.

We then define the ***Required Performance*** of the measured performance characteristic to be the $(1-p) \cdot 100^{th}$ percentile, Q_{1-p}, for a lower requirement or the $p \cdot 100^{th}$ percentile, Q_p, for an upper requirement. This implies that these percentiles relate directly to the maximum probability of failure. That is, for a lower requirement, $Prob\left(PC < Q_{1-p}\right) = 1 - p = P_{req}$, and for an upper requirement, $Prob\left(PC > Q_p\right) = 1 - Prob\left(PC < Q_p\right) = 1 - p = P_{req}$. Figure 4.1 below depicts these percentiles and their interpretation graphically.

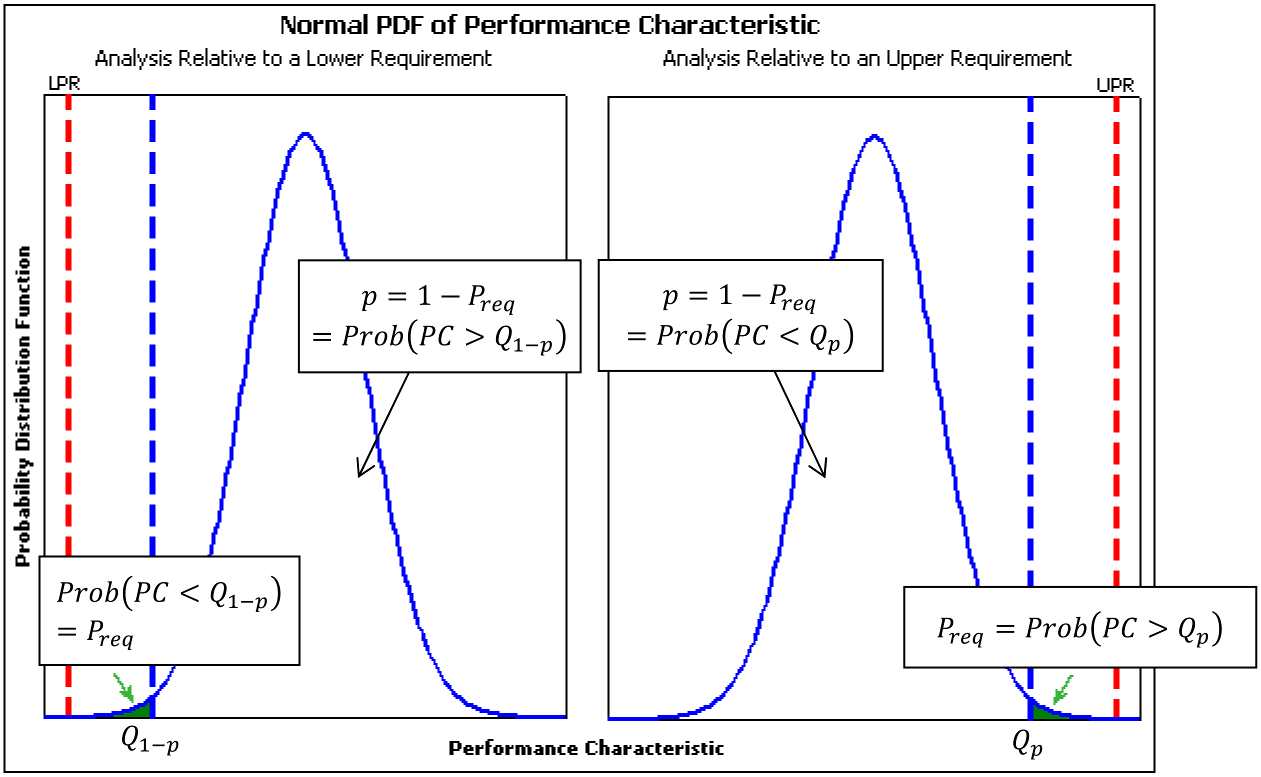

Figure 4.1. Graphical Depiction of the Percentile, the Content, and P_{req}.

In general, we will not know the distribution of the performance characteristic perfectly, but rather will estimate it from a sample of data. We define the ***Observed Performance*** or ***Assessed Performance*** as the ***Estimate*** of the chosen percentile and denote this by either $\hat{Q}_{1-p}$ or $\hat{Q}_p$ for a lower or upper percentile respectively.

The observed performance is an estimate of a population parameter based on a sample of data. To account for the sampling uncertainty, the range of possible values of the population parameter may be computed from the sample with a defined likelihood (a statistical confidence level) of bounding the true population parameter with a ***Confidence Bound***. Suppose we desire a $\gamma \cdot 100\%$ confidence bound on the estimated percentile. For a lower percentile, $\hat{Q}_{1-p}$, a lower confidence bound is computed, denoted by $\hat{Q}_{1-p,\gamma}$, and for an upper percentile, $\hat{Q}_p$, an upper confidence bound is computed, denoted by $\hat{Q}_{p,\gamma}$. The confidence bound accounts for the uncertainty in the estimation of the desired percentile and is referred to as a ***Statistical Tolerance Bound*** [5]. Formally,

$$Prob\left(Prob\left(PC < \hat{Q}_{1-p,\gamma}\right) \le 1-p\right) \ge \gamma \quad \text{or} \quad Prob\left(Prob\left(PC < \hat{Q}_{p,\gamma}\right) \ge p\right) \ge \gamma.$$

As discussed in Sections 1, the tolerance interval covers a specified proportion of a population with a certain confidence level, which is exactly the goal of the QMU questions posed above. For an analysis with respect to a lower requirement (LPR), this tolerance bound, $\hat{Q}_{1-p,\gamma}$, is the value that $p \cdot 100\%$ of the performance characteristic values (the content) will be greater than with $\gamma \cdot 100\%$ confidence. Therefore, if $\hat{Q}_{1-p,\gamma} > LPR$ then we are able to claim that $p \cdot 100\%$ of the performance characteristic values will be greater than the lower performance requirement with $\gamma \cdot 100\%$ confidence. This also implies that if $\hat{Q}_{1-p,\gamma} > LPR$ then we can claim that at most $(1-p) \cdot 100\% = P_{req} \cdot 100\%$ of the performance characteristic values will be less than the lower performance requirement with $\gamma \cdot 100\%$

confidence. A similar statement could be made for an upper requirement however we would require $\hat{Q}_{p,\gamma} < UPR$. Therefore, the tolerance bound incorporates information about margin and uncertainty and can be compared directly to the performance requirement to draw conclusions. This is appealing because all decisions remain on the engineering unit scale, which provides an easily interpreted result.

We also introduce the concept of a ***Coverage Probability***. The coverage probability is defined as the estimated probability that a performance characteristic will be either greater than a lower requirement or less than an upper requirement with $\gamma \cdot 100\%$ confidence. That is, the coverage probability is the value p_c that satisfies the equation, $\hat{Q}_{1-p_c,\gamma} = LPR$ for a lower requirement or $\hat{Q}_{p_c,\gamma} = UPR$ for an upper requirement. In other words, it is the content value such that the tolerance bound will be exactly equal to the performance requirement. The coverage probability can be calculated by iteratively changing the content value until the tolerance bound equals the performance requirement (easily implemented by a linear search algorithm). This metric will be useful in exploring the properties of the new figure-of-merit, which is defined in Section 4.2. The following discussion provides explicit definitions for the computation of the margin and uncertainty.

If the estimated lower percentile is greater than the lower performance requirement (LPR), we interpret this as demonstrating that there is positive margin to the lower performance requirement. Similarly, if the estimated upper percentile is less than the upper performance requirement (UPR) then this demonstrates positive margin to the upper performance requirement. Formally, the ***Margin*** is defined to be the difference between the percentile and the performance requirement, $M = Q_{1-p} - LPR$, for a lower requirement or, $M = UPR - Q_p$, for an upper requirement. Similarly, the ***Estimated Margin*** is defined to be the difference between the estimated percentile and the performance requirement, $\widehat{M} = \hat{Q}_{1-p} - LPR$, for a lower requirement or, $\widehat{M} = UPR - \hat{Q}_p$, for an upper requirement. As mentioned above, in some applications the appropriate metric could be defined as the median Q_{50} or as the mean (as in the k-factor methodology) of the performance characteristic distribution. In such cases, the definition of margin would still hold as defined here.

Next, we define ***Uncertainty*** as the width of the confidence bound (absolute difference between the estimated percentile and its confidence bound), $\widehat{U} = \hat{Q}_{1-p} - \hat{Q}_{1-p,\gamma}$ for a lower requirement, and $\widehat{U} = \hat{Q}_{p,\gamma} - \hat{Q}_p$ for an upper requirement This definition accounts for the uncertainty in the estimation of the desired percentile, which in most cases is a function of both the mean and standard deviation. Therefore, we ultimately account for the uncertainty in the estimation of both of these parameters rather than just the mean.

It should be noted that this definition of uncertainty only accounts for the sampling uncertainty in the estimation of the chosen percentile. The confidence bound quantifies the range of potential values for the true population parameter around the best estimate of that parameter. Although the uncertainty defined here only accounts for sampling variability, the overall methodology still quantifies the aleatory uncertainty from stochastic variability with the definition of margin above. Once a probability distribution function is specified, the stochastic variability is captured by the shape of the assumed distribution. Therefore, the estimation of the desired percentile, which requires a distributional assumption, already quantifies the unit-to-unit variability. Hence, the estimate of margin defined here captures the uncertainty from unit-to-unit variability (in addition to information with respect to meeting the defined performance requirements) and the estimate of uncertainty defined here captures the uncertainty from sampling variability, both of which are forms of aleatory uncertainty.

4.2. A New Figure of Merit

We define a new *figure-of-merit* to be a ***Tolerance Ratio*** ($\boldsymbol{TR}$)**,** which is the ratio of estimated margin divided by the estimated uncertainty based on the tolerance bound methodology,

$$TR = \frac{\hat{M}}{\hat{U}} = \frac{\hat{Q}_{1-p} - LPR}{\hat{Q}_{1-p} - \hat{Q}_{1-p,\gamma}} \quad \text{or} \quad TR = \frac{UPR - \hat{Q}_p}{\hat{Q}_{p,\gamma} - \hat{Q}_p}.$$

This figure-of-merit definition incorporates the performance requirement, the maximum allowable probability of failure, the statistical confidence level, and the aleatory uncertainty arising from both unit-to-unit variability and sampling variability. It does not however account for issues with measurement bias or instabilities in the measurement process over time. If the accuracy of the measurement system is flawed, causing a systematic shift or bias to all measurements, then the tolerance ratio could provide misleading results. Therefore, it is assumed that the measurement uncertainty is thoroughly explored during a comprehensive engineering analysis as discussed in Section 5.

Recall the decision criteria discussed in Section 4.1 above consisted of a comparison of the tolerance bound and the performance requirement. To remain consistent with existing methodologies and make decisions based on the margin divided by uncertainty figure-of-merit, we have, if $\hat{Q}_{1-p,\gamma} > LPR$ then $\hat{Q}_{1-p} - \hat{Q}_{1-p,\gamma} < \hat{Q}_{1-p} - LPR$ indicates we are meeting the requirement, and hence

$$TR = \frac{\hat{Q}_{1-p} - LPR}{\hat{Q}_{1-p} - \hat{Q}_{1-p,\gamma}} > 1$$

shows we are meeting the requirements as well. Therefore, comparing the new tolerance ratio against the critical value of 1 is the only decision criteria needed. We do acknowledge that the computation of this metric still requires a decision on the maximum probability of failure to demonstrate and the statistical confidence level, both of which should be meaningful and documented. The choice of these values should be decided in the engineering analysis of the data prior to any formal analyses. We believe it is more natural to have the estimate of the figure-of-merit, TR, change based on changes in these two values rather than having the critical value or decision rule change. The Sections 4.3 and 4.4 provide additional details of these new methodologies specific to a point-in-time or regression analysis respectively.

4.3. Point in Time Analysis

Recall, a point-in-time analysis is performed if data are collected at a single point in time, such as at product acceptance, or if no trend is present in the data. The goal of this analysis is to quantify the margin and uncertainty of the measured performance characteristic (PC) relative to an upper or lower performance requirement (UPR or LPR) at that point in time. A point in time analysis attempts to answer the following question.

1. Are we $\boldsymbol{YY}$% certain that **at-least** $\boldsymbol{XX}$% of the unit population will yield a response **greater than** the threshold $\boldsymbol{T}$?

Further recall, the values of $\boldsymbol{XX}$, $\boldsymbol{YY}$, $\boldsymbol{Z}$, and $\boldsymbol{T}$ and the comparisons '**at-least**' versus '**at-most**' and '**greater than**' versus '**less than**' are all parameters of the requirement and will be specific to each individual analysis.

The estimated margin is defined as the difference between the estimate of the chosen percentile and the performance requirement, $\widehat{M} = \hat{Q}_{1-p} - LPR$, for a lower requirement. The uncertainty is defined as the absolute difference between the estimated percentile and its confidence bound, $\widehat{U} = \hat{Q}_{1-p} - \hat{Q}_{1-p,\gamma}$. Figure 4.2 depicts these new definitions using a Normal distribution. The metrics are defined in the figures and equations below relative to the lower performance requirement (LPR) for reference. For an upper bound, the metrics could be adjusted appropriately.

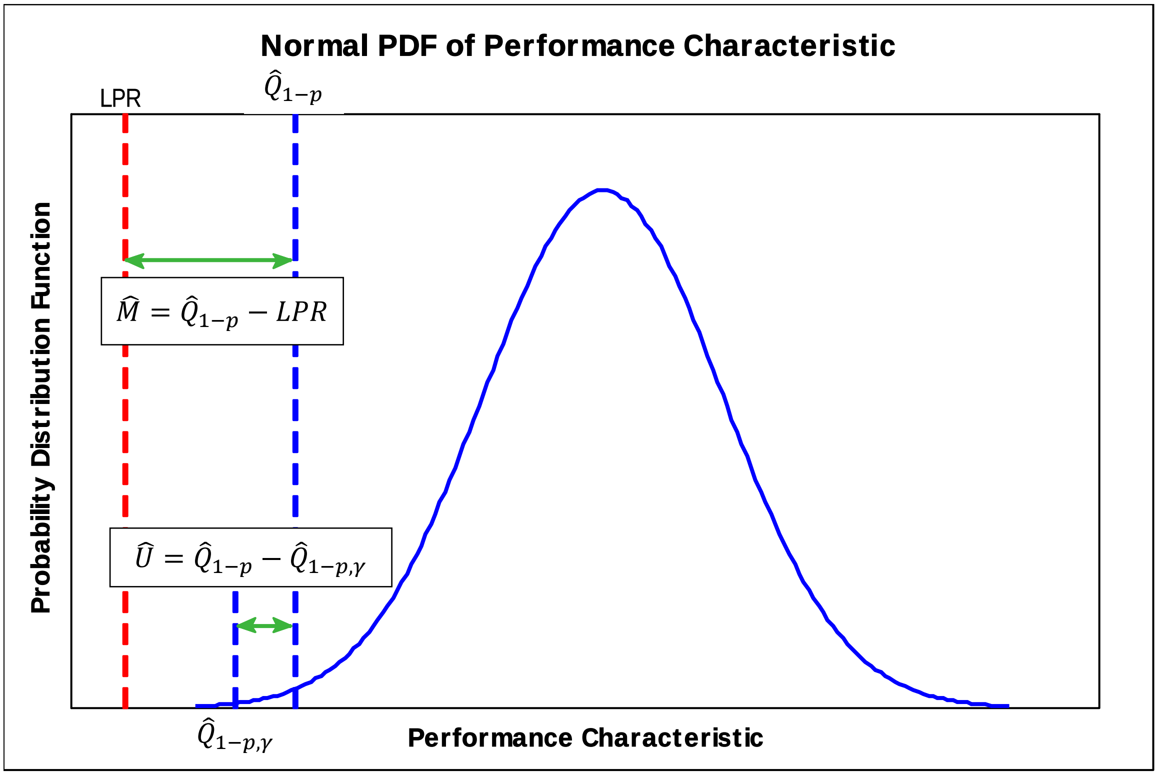

Figure 4.2. Graphical Depiction of the New Metrics Relative to a Normally Distributed PC.

Although this new methodology does require a distributional assumption, the choice of which should be a key part of the analysis process, it can be applied to any standard distribution. Figure 4.3 demonstrates this by depicting these metrics using a Weibull distribution fit. While the estimation of these metrics will be dependent on the chosen distribution, the definition of the metrics, their meaning and interpretation, and most importantly the decision criteria and critical value remain the same regardless of the choice of distribution. This provides a consistent and interpretable methodology that can be applied across a wide variety of datasets.

The implementation of this methodology and the estimation of the tolerance bounds when the data do not follow a Normal distribution may be difficult, especially for a non-statistician. There are two substantial practical hurdles. First, one must choose the best distributional form to use for the given dataset and next one must be able to correctly calculate the tolerance bound for the chosen distribution. Both steps add complexity to the QMU process, however, these steps need only to be applied to a select few performance characteristics that are mature enough and have been through a rigorous engineering analysis. Once a dataset arrives at this stage of the QMU process, the component engineers and/or analysts should consult with a statistician to decide on the best statistical approach. Widespread adoption and implementation of these methodologies can be achieved through collaboration.

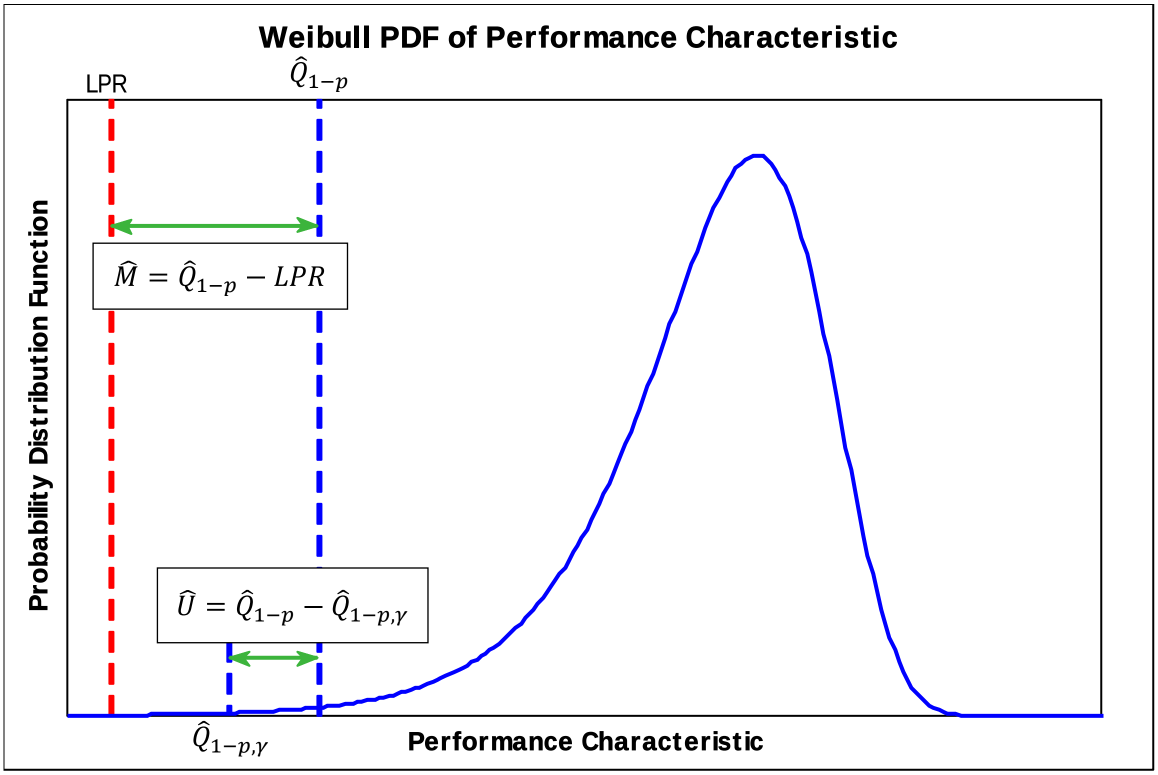

Figure 4.3. Graphical Depiction of the New Metrics Relative to a Weibull Distributed PC.

The remainder of this section provides the technical details for the implementation of this point-in-time methodology. Below we show examples using the Normal distribution (Example Dataset 1), the Weibull distribution (Example Dataset 3), and Lognormal distribution (Example Dataset 4). For more details on the implementation of this methodology for other distributions, please refer to reference 5.

4.3.1 Normal Data Analysis

We begin by analyzing Example Dataset 1 using a Normal distribution fit. For a univariate Normal distribution with mean μ and standard deviation σ, the $r \cdot 100^{\text{th}}$ percentile is,

$$Q_r = \mu + \sigma \cdot \Phi^{-1}(r),$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) for a standard Normal distribution ($\mu = 0$, $\sigma = 1$). That is, $\Phi^{-1}(r)$ is the $r \cdot 100^{\text{th}}$ percentile of a standard Normal distribution. Given a sample of data the mean and standard deviation can be estimated by the sample mean, $\bar{x}$, and sample standard deviation, s. For a sample of size n these are defined as,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The best estimate of this percentile is obtained by replacing the mean and standard deviation with their respective best estimates, $\bar{x}$ and s. Hence,

$$\hat{Q}_r = \bar{x} + s \cdot \Phi^{-1}(r).$$

For Example Dataset 1, shown in Figure 4.4 below, there are 65 observations yielding an estimated mean and standard deviation of $\bar{x} = 9.993$ and $s = 0.241$. For this dataset there is a lower performance requirement of $LPR = 9$. Further, for example, assume we require $Prob(PC < LPR) < P_{req} = 0.005$ with 95% confidence. Hence, the content is $p = 1 - P_{req} = 0.995$. To frame this into the structure of question introduced above, we have

1. Are we **95%** certain that **at-least 99.5%** of the unit population will yield a response **greater than** the threshold **$LPR = 9$**?

Since the analysis is with respect to a lower bound the desired percentile would be the $(1 - p) \cdot 100 = 0.5^{\text{th}}$ percentile, $Q_{1-p} = Q_{1-0.995} = Q_{0.005}$, and $\Phi^{-1}(0.005) = -2.576$. Therefore, the estimated 0.5^{th} percentile for this set of data is,

$$\hat{Q}_{0.005} = 9.993 + 0.241 \cdot \Phi^{-1}(0.005) = 9.372.$$

For data that follows a Normal distribution the estimated lower tolerance bound from a sample of size n is of the form,

$$\hat{Q}_{1-p,\gamma} = \bar{x} - s \cdot k_1,$$

where

$$k_1 = t_{n-1,\gamma}\left(\sqrt{n} \cdot \Phi^{-1}(p)\right)/\sqrt{n}$$

and $t_{df,\gamma}(\Delta)$ denotes the $\gamma \cdot 100^{\text{th}}$ percentile of a non-central t-distribution with df degrees of freedom and noncentrality parameter Δ. The value k_1 is commonly referred to as a tolerance factor and should not be confused with the k-factor discussed in Section 2. For details on the derivation of this tolerance factor and the tolerance bound in general, please refer to reference 5. Tables of the non-central t-distribution are also available in a Sandia Monograph on tolerance intervals [6]. An upper tolerance bound for a Normal distribution, computed when the analysis is with respect to an upper limit, is given by, $\hat{Q}_{p,\gamma} = \bar{x} + s \cdot k_1$.

For Example Dataset 1,

$$k_1 = t_{64,0.95}\left(\sqrt{65} \cdot \Phi^{-1}(0.995)\right)/\sqrt{65} = 3.072$$

and

$$\hat{Q}_{0.005,0.95} = 9.993 - 0.241 \cdot 3.072 = 9.253.$$

The estimated tolerance bound is greater than the lower performance requirement, which indicates that we are meeting the requirement. Therefore, we can assert that we are 95% confident that at least 99.5% of the performance characteristic values will be greater than the lower performance requirement of 9. The following show explicit calculations of the margin, uncertainty, and tolerance ratio and show how the same conclusion can be reached via those metrics.

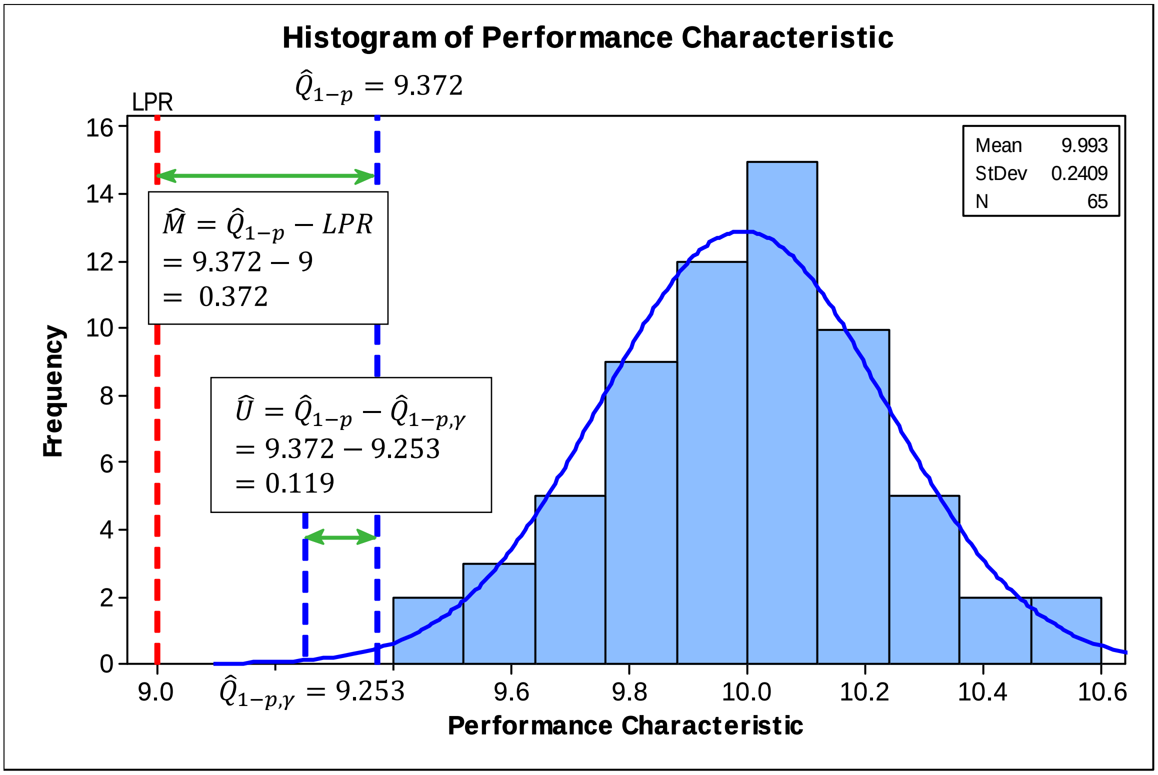

Figure 4.4. Example Dataset 1 Showing the Metrics for the New QMU Methodology.

The estimated margin and uncertainty for this dataset is, $\hat{M} = \hat{Q}_{1-p} - LPR = 9.372 - 9 = 0.372$ and $\hat{U} = \hat{Q}_{0.005} - \hat{Q}_{0.005,0.95} = 9.372 - 9.253 = 0.119$. A depiction of these new metrics applied to Example Dataset 1 is shown in Figure 4.4 above. Finally, the estimate of the tolerance ratio is given by,

$$TR = \frac{\hat{M}}{\hat{U}} = \frac{0.372}{0.119} = 3.13.$$

Since $TR = 3.13 > 1$, we can conclude that we are meeting the requirements and can assert that we are 95% confident that at least 99.5% of the performance characteristic values will be greater than the lower performance requirement of 9. Again, this should be obvious since the estimated tolerance bound, $\hat{Q}_{0.005,0.95} = 9.253$, is greater than the lower performance requirement, which indicates that we are meeting the requirement. Further, for a fixed confidence level of 95%, a tolerance bound with a content of $p = 0.99975$ is estimated to be, $\hat{Q}_{0.99975\,,0.95} = 9$. Therefore, the coverage probability is estimated to be, $p_c = 0.99975$. The magnitude of the tolerance ratio indicated that we could actually make a stronger statement, which we have now quantified with the computation of the coverage probability. We can assert that, with 95% confidence, at least 99.975% of the performance characteristic values will be greater than the lower performance requirement of 9. We also note that this conclusion is consistent with the conclusion based on the k-factor shown in Section 2.1. This however should be expected for data that follows a Normal distribution. We will see, using the non-normal Example Dataset 3, that this will not always be the case.

One main advantage of this new methodology is the interpretability. If the lower tolerance bound is greater than the lower performance requirement then we are meeting our requirements and we will have $TR > 1$. If we change our requirements, to say $P_{req} = 0.0001$ with 95% confidence, then we would compute a new tolerance ratio but the decision criteria will remain the same. For Example Dataset 1, consider, $p = 1 - P_{req} = 0.9999$. The estimated percentile and confidence bound are, $\hat{Q}_{1-p} = \hat{Q}_{0.0001} = 9.993 + 0.241 \cdot \Phi^{-1}(0.0001) = 9.097$ and $\hat{Q}_{1-p,\gamma} = \hat{Q}_{0.0001,0.95} = 8.933$. The estimated margin and uncertainty are, $\hat{M} = \hat{Q}_{1-p} - LPR = 9.097 - 9 = 0.097$, and $\hat{U} = \hat{Q}_{0.0001} - \hat{Q}_{0.0001,0.95} = 9.097 - 8.933 = 0.1637$. Clearly, the tolerance bound $\hat{Q}_{0.0001,0.95} = 8.933$ is less than the lower performance requirement so we are not meeting the requirement and necessarily $TR = 0.593$ is less than 1. We have changed the requirement level, producing a new estimate of the figure-of-merit, but still have the same decision criteria and critical value. Therefore, this dataset does not meet the requirement of $P_{req} = 0.0001$ with 95% confidence.

4.3.2 Non-Normal Data Analysis

This section presents methodologies for performing a point in time analysis on non-Normal data. There are two different approaches that will be discussed, a transformation approach and a direct parametric approach using a standard but non-Normal distribution. The transformation approach is appropriate when the data can be transformed easily (with a one-to-one transformation) to a Normal distribution, for example with a log or square root transformation. If the data fits a non-Normal distribution that does not have a one-to-one transformation with a Normal distribution (e.g. a Weibull distribution) then the direct parametric approach will be more appropriate.

Transformation Approach

This section describes an approach for non-Normal data when there is a one-to-one transformation available to transform the observed data to fit a Normal distribution. Consider Example Dataset 4 in Figure 4.5 with a Lognormal fit to the data. For this dataset suppose there is a lower performance requirement of $LPR = 0.25$.

If the performance characteristic distribution fits a Lognormal distribution then $\ln(PC)$ follows a Normal distribution. Suppose, $y_1, y_2, \ldots, y_n$ are a sample from a Lognormal performance characteristic distribution with location and scale parameters μ and σ respectively. Then, $x_1 = \ln(y_1), x_2 = \ln(y_2), \ldots, x_n = \ln(y_n)$ is a sample from a Normal distribution with mean and standard deviation μ and σ respectively. Therefore, the Normal based approaches discussed in Section 4.3.1 can be applied to construct tolerance bounds based on the sample $x_1, x_2, \ldots, x_n$. Recall, $\hat{Q}_{1-p,\gamma} = \bar{x} - s \cdot k_1$ and $\hat{Q}_{p,\gamma} = \bar{x} + s \cdot k_1$ are lower and upper tolerance bounds respectively for a sample from a Normal distribution where,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad k_1 = \frac{t_{n-1,\gamma}\left(\sqrt{n} \cdot \Phi^{-1}(p)\right)}{\sqrt{n}},$$

and $t_{df,\gamma}(\Delta)$ denotes the $\gamma \cdot 100^{\text{th}}$ percentile of a non-central t-distribution with df degrees of freedom and noncentrality parameter Δ. Therefore,

$$\hat{Q}^*_{1-p,\gamma} = e^{\hat{Q}_{1-p,\gamma}} = e^{\bar{x} - s \cdot k_1} \text{ and } \hat{Q}^*_{p,\gamma} = e^{\hat{Q}_{p,\gamma}} = e^{\bar{x} + s \cdot k_1}$$

are lower and upper tolerance bounds respectively from the original Lognormal performance characteristic distribution. Furthermore, $Prob(PC < LPR) = Prob(\ln(PC) < \ln(LPR))$, therefore all questions posed in Section 4.3.1 based on the Normal distribution can be readily applied.

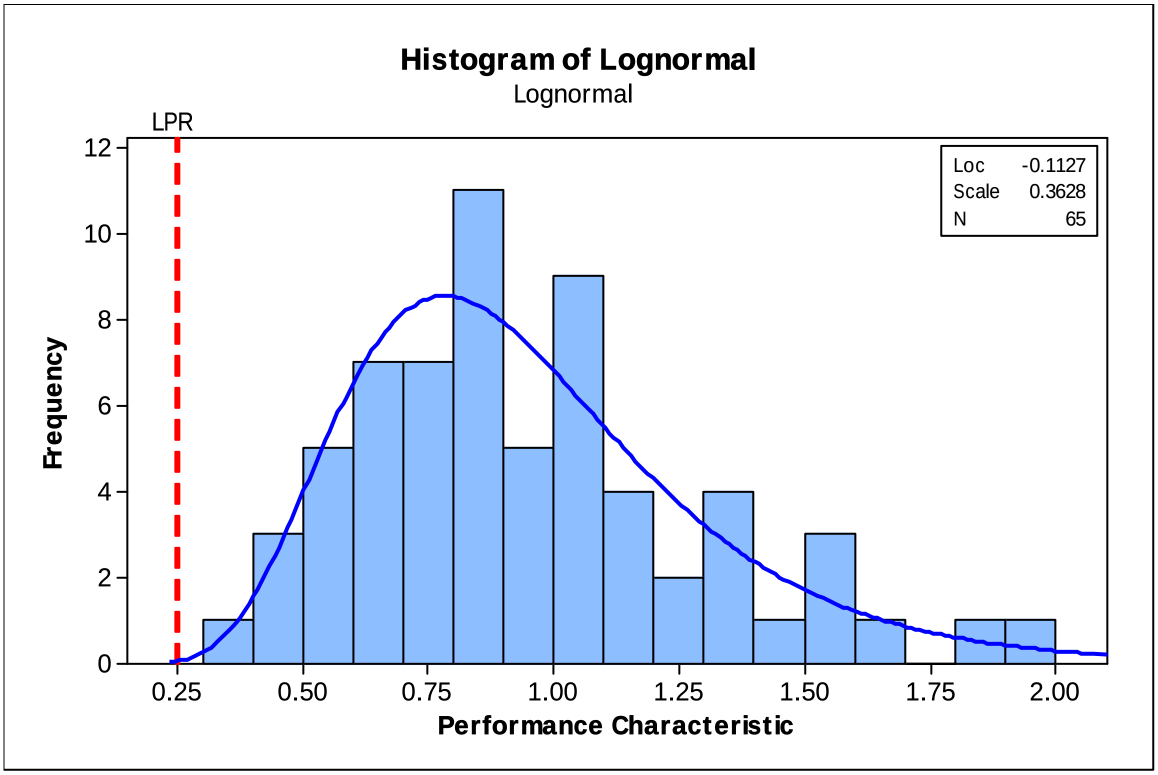

Figure 4.5. Example Dataset 4 with a Lognormal Distribution Fit.

For Example Dataset 4, the mean and standard deviation of the transformed Normal data are equivalent to the location and scale parameters from the original Lognormal distribution which are shown in Figure 4.5. Hence, $\bar{x} = -0.113$ and $s = 0.363$ are obtained from a sample of size $n = 65$. For example, assume there is a requirement that 99% of the units must have a performance characteristic value greater than the lower performance requirement with 95% confidence. The formal QMU question is then,

1. Are we **95%** certain that **at-least 99%** of the unit population will yield a response **greater than** the threshold $\boldsymbol{LPR = 0.25}$?

The question posed requires $Prob(PC < LPR) < P_{req} = 0.01$ with 90% confidence. Hence, the content for this analysis is $p = 1 - P_{req} = 0.99$ and the desired percentile would be the $(1 - p) \cdot 100 = 1^{st}$ percentile, $Q_{1-p} = Q_{0.01}$, and $\Phi^{-1}(0.01) = -2.326$. Therefore, the estimated 1^{st} percentile for transformed set of data is,

$$\hat{Q}_{0.01} = -0.113 + 0.363 \cdot \Phi^{-1}(0.01) = -0.957.$$

The estimated 1^{st} percentile for the original Lognormal set of data is,

$$\hat{Q}^{*}_{0.01} = e^{\hat{Q}_{0.01}} = e^{-0.957} = 0.384.$$

As discussed above, the confidence bound on the estimated percentile is a statistical tolerance bound that can be computed using the Normal distribution methodology. For this dataset, $n = 65$, $df = n - 1 = 64$, $\Delta = \sqrt{65} \cdot \Phi^{-1}(0.99) = 18.76$, and

$$k_1 = t_{64,0.95}(18.76)/\sqrt{65} = 2.79.$$

The lower tolerance bound for the transformed dataset is estimated to be

$$\hat{Q}_{0.01,0.95} = -0.113 - 0.363 \cdot 2.79 = -1.124$$

and the estimated lower tolerance bound for the original Lognormal set of data is,

$$\hat{Q}^*_{0.01,0.95} = e^{\hat{Q}_{0.01,0.95}} = 0.325.$$

The estimated lower tolerance bound, $\hat{Q}_{0.01,0.95} = 0.325$, is greater than the lower performance requirement, which indicates that we are meeting the requirement. We reinforce this conclusion by calculating the formal figure-of-merit. The estimated lower margin for this dataset is, $\hat{M} = \hat{Q}^*_{0.01} - LPR = 0.384 - 0.25 = 0.134$ and the estimated uncertainty is $\hat{U} = \hat{Q}^*_{0.01} - \hat{Q}^*_{0.01,0.95} = 0.384 - 0.325 = 0.059$. Therefore, the estimated tolerance ratio is given by,

$$TR = \frac{\hat{M}}{\hat{U}} = \frac{0.134}{0.059} = 2.27.$$

Since $TR = 2.27 > 1$, we can conclude that we are meeting the requirements and can assert that we are 95% confident that at least 99% of the performance characteristic values will be greater than the lower performance requirement of 0.25. The estimates for this analysis are depicted in Figure 4.6 below.

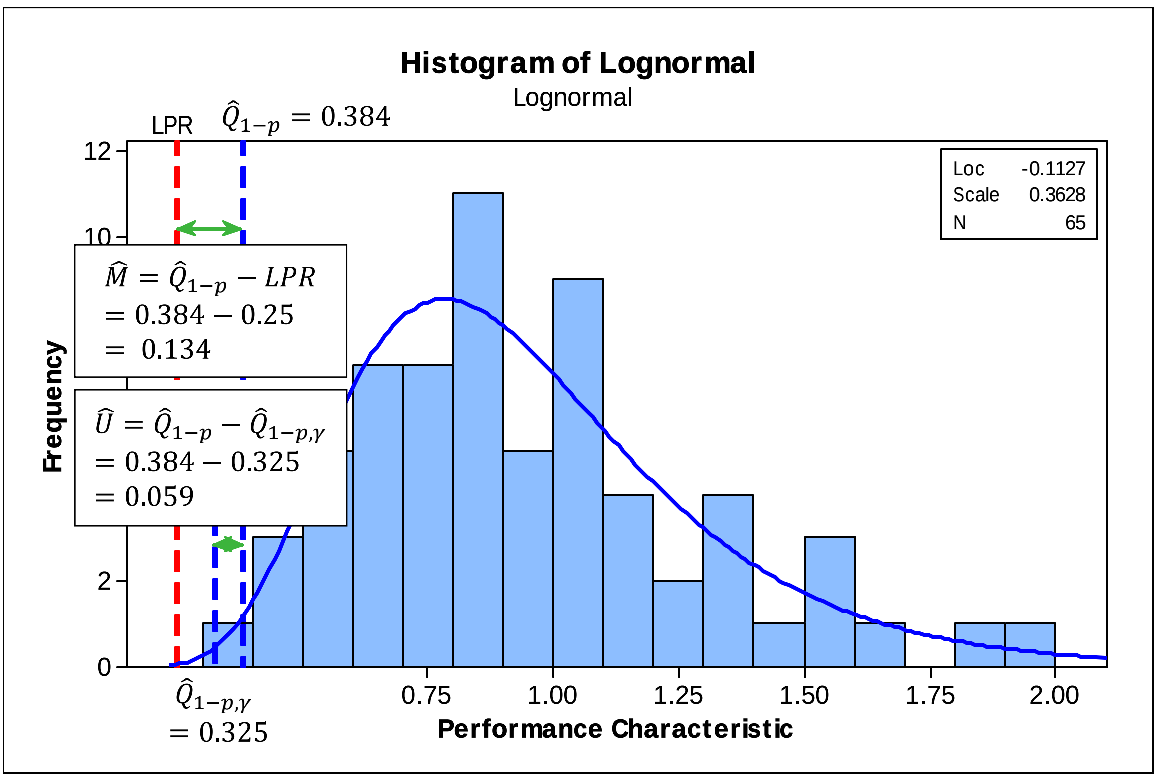

Figure 4.6. QMU Metrics for Example Dataset 2 with a Lognormal Distribution Fit.

Again, note that these metrics are shown on the engineering unit scale. Although a transformation was required for their computations, the constraint that the transformation must be a one-to-one transformation allows the metrics to be translated easily back to their original scale, as shown in the equations and figure above. The following discussion shows an example when a one-to-one transformation is not available.

Direct Parametric Approach

Next, we analyze Example Dataset 3 using a Weibull distribution fit. Recall, from the probability plots in Figure 3.3 above the Weibull distribution appeared to be an adequate assumption (p-value = 0.137) for this dataset. The two-parameter Weibull distribution is characterized by a shape (β) parameter and a scale (η) parameter. The Weibull distribution can take on several different parameterizations. Here we use the PDF form given by,

$$f(x) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^{\beta}}.$$

The Weibull distribution does not have a one-to-one transformation with the Normal distribution. For a two-parameter Weibull distribution with shape β and scale η, the $r \cdot 100^{\text{th}}$ percentile is,

$$Q_r = \eta \cdot [-\ln(1-r)]^{1/\beta}.$$

The best estimate of this percentile is given by replacing the shape and scale parameters with their respective maximum likelihood estimates, $\hat{\beta}$ and $\hat{\eta}$. Hence,

$$\hat{Q}_r = \hat{\eta} \cdot [-\ln(1-r)]^{1/\hat{\beta}}.$$

For Example Dataset 3, shown again in Figure 4.7, there are 65 observations yielding an estimated shape and scale of $\hat{\beta} = 51.49$ and $\hat{\eta} = 10.14$. To be consistent with the example used previously we will again assume we require $P < P_{req} = 0.005$ or the desired content is $p = 1 - P_{req} = 0.995$.

1. Are we $\mathbf{95\%}$ certain that **at-least** $\mathbf{99.5\%}$ of the unit population will yield a response **greater than** the threshold $\boldsymbol{LPR = 9}$?

Since we are estimating a lower bound, the desired percentile would be the 0.5$^{\text{th}}$ percentile ($1 - p = 0.005$). Therefore, the estimated 0.5$^{\text{th}}$ percentile for this set of data is,

$$\hat{Q}_{0.005} = 10.14 \cdot [-\ln(1-0.005)]^{\frac{1}{51.49}} = 9.149.$$

For data that follows a Weibull distribution the estimated lower tolerance bound from a sample of size n is of the form,

$$\hat{Q}_{1-p,\gamma} = \hat{\eta} \cdot \exp\left(w_{1-p,1-\gamma}/\hat{\beta}\right),$$

where $w_{1-p,1-\gamma}$ is the $(1-\gamma) \cdot 100^{\text{th}}$ percentile of $w = \beta^*[-\ln(\eta^*) + \ln(-\ln(p))]$ and β^* and η^* are the maximum likelihood estimates calculated from a sample of size n from a Weibull ($\beta = 1, \eta = 1$) distribution. The distribution of w does not depend on any unknown parameters, and so its percentiles can be estimated using Monte Carlo simulation. For details on the derivation of this tolerance bound, refer to reference 5. For an upper tolerance bound, $\hat{Q}_{p,\gamma} = \hat{\eta} \cdot \exp\left(w_{p,\gamma}/\hat{\beta}\right)$ where $w_{p,\gamma}$ is the $\gamma \cdot 100^{\text{th}}$ percentile of $w = \beta^*[-\ln(\eta^*) + \ln(-\ln(1-p))]$.

For Example Dataset 3, $w_{0.005,1-0.95} = -6.43$, based on a Monte Carlo simulation consisting of 100,000 runs, and

$$\hat{Q}_{0.005,0.95} = 10.14 \cdot \exp\left(\frac{-6.43}{51.49}\right) = 8.95,$$

which is less than the lower performance requirement. Consequently, we can conclude that we are **not** meeting the requirements and **cannot** assert that we are 95% confident that at least 99.5% of the performance characteristic values will be greater than the lower performance requirement of 9. The margin, uncertainty, and tolerance ratio are calculated exactly as in the Normal case. This feature enhances the interpretability of this methodology.

The estimated margin for this dataset is, $\widehat{M} = \hat{Q}_{1-p} - LPR = 9.149 - 9 = 0.149$. Further, the estimated uncertainty is, $\widehat{U} = \hat{Q}_{0.005} - \hat{Q}_{0.005,0.95} = 9.149 - 8.95 = 0.199$, and the estimate of the new figure-of-merit is given by,

$$TR = \frac{\widehat{M}}{\widehat{U}} = \frac{0.149}{0.199} = 0.75,$$

which is less than 1 reinforcing the fact that, based on the available data, we are not meeting the requirements. A depiction of these new metrics applied to Example Dataset 3 is shown in Figure 4.7 below.

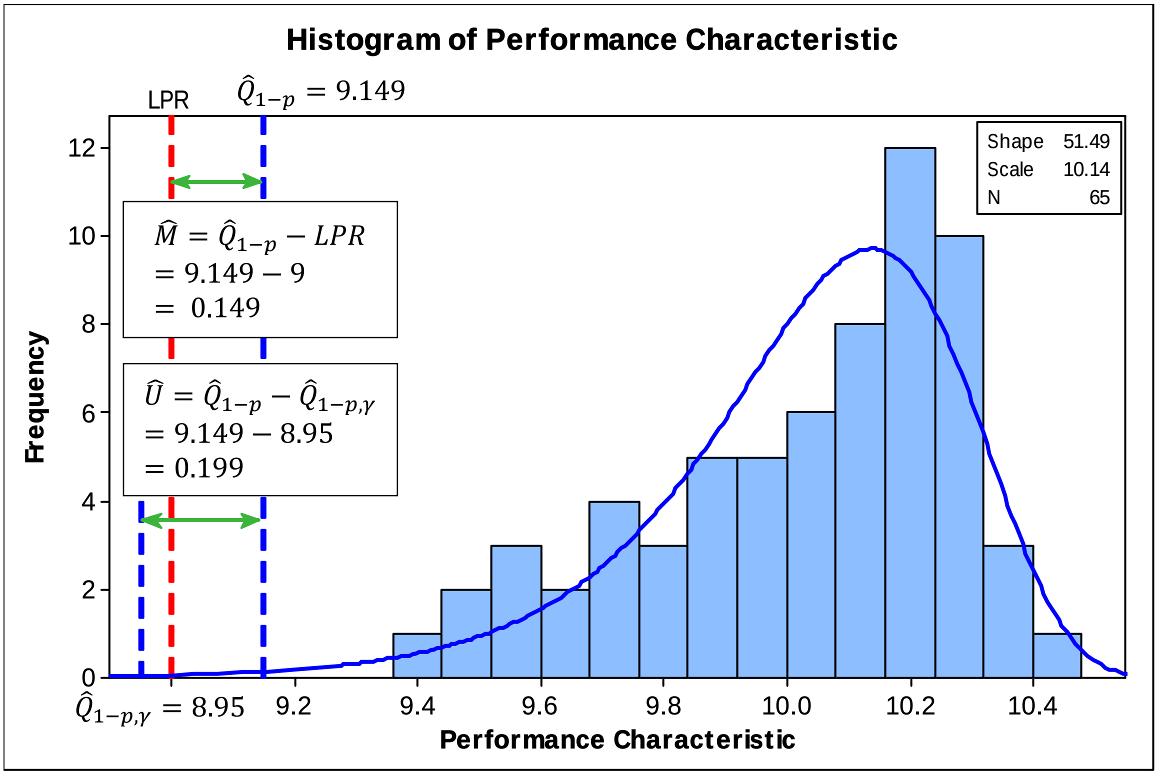

Figure 4.7. Example Dataset 3 Showing the Metrics for the New QMU Methodology.

As mentioned previously, when we fail to meet the specified requirements, a coverage probability can be computed to explore the true probability of meeting the requirements with a fixed level of statistical confidence. For a fixed confidence level of 95%, a tolerance bound with a content of $p = 0.9935$ is estimated to be, $\hat{Q}_{0.9935\,,0.95} = 9$. Hence, the coverage probability is estimated to be, $p_c = 0.9935$. The tolerance ratio provides the initial conclusion that we cannot assert that 99.5% of the performance characteristic values are greater than the lower performance requirement with 95% confidence. The computation of the coverage probability then allows us to quantify the degree to which we fail to meet this requirement. Here, we can only claim that, with 95% confidence, at most 99.35% of the performance characteristic values will be greater than the lower performance requirement.

Recall, the k-factor analysis of Example Dataset 3 resulted in a sample mean and sample standard deviation of $\bar{x} = 10.027$ and $s = 0.257$. Thus, the estimated k-factor for these data, relative to a lower performance requirement of 9 was, $\hat{k} = 3.99$ and the 95% lower confidence bound on the k-factor was estimated to be $\hat{k}_{0.95} = 3.38$ which indicates good margin. In fact, comparing $\hat{k}_{0.95}$ to a critical value of 2.576 for a maximum probability of failure of 0.005 we would conclude that we have met this requirement. The mean and standard deviation alone however do not accurately characterize the shape of this distribution. By properly characterizing the distribution one can see that the tail of this Weibull distribution fit (Figure 4.7) is much wider than a Normal distribution fit for the same dataset (Figure 3.2).

This demonstrates the importance of properly characterizing the distribution of the performance characteristic to determine if it meets requirements. Once that is accomplished, the estimation of a percentile of that distribution (and its confidence bound) provides the information needed to make informed decisions. The next section demonstrates how this methodology can be applied for a dataset with an observed aging trend.

4.4. Age Trend Regression with a Tolerance Bound

This section presents the technical details to obtain an estimated percentile of a regression fit and a confidence bound on that percentile (statistical tolerance bound). The identification of the age trend, along with a determination that the trend is due to an aging effect and not some other known or unknown factor, must be done during the engineering analysis of the dataset. When an age trend is present in the data, the goal of the analysis tends to attempt to answer one of the following questions,

1. Are we $\boldsymbol{YY}$% certain that **at-least** $\boldsymbol{XX}$% of the unit population will yield a response **greater than** the threshold $\boldsymbol{T}$ after $\boldsymbol{Z}$ years of life?

2. At which age will we no longer be $\boldsymbol{YY}$% certain that **at-least** $\boldsymbol{XX}$% of the unit population will yield a response **greater than** the threshold $\boldsymbol{T}$?

This remainder of this section describes the statistical analysis needed to estimate an ***Alarm Age***. The alarm age is defined as the component age at which we estimate certain percentage of the population is no longer contained by the performance requirement, with a given level of confidence. Measured values of a performance characteristic may trend up or down, or the range of values may grow larger or smaller as components age. We are most concerned if the measured values trend toward a performance requirement or if the range of measured values expands toward the limit. Figure 4.8 shows a notional plot of the mean performance characteristic versus component age (CA). Here the mean, $\mu(CA)$, the percentiles, $Q_{1-p}(CA)$, the margin, $M(CA) = Q_{1-p}(CA) - LPR$, and uncertainty, $U(CA) = Q_{1-p}(CA) - Q_{1-p,\gamma}(CA)$, are all functions of the component age. The notional data in Figure 4.8 depicts a downward trend and increasing measurement error bars with increasing component age. A lower performance requirement (LPR) is shown by the red dashed line. The plot thus demonstrates an example of decreasing performance with increasing variability, with possible failure at some point in the future. Again, the case of increasing variability with age is shown here for example purposes only. The regression analyses described throughout this paper assume a constant variance. For data that exhibits increasing variability with age, a statistician should be consulted to determine the appropriate methodology to apply.

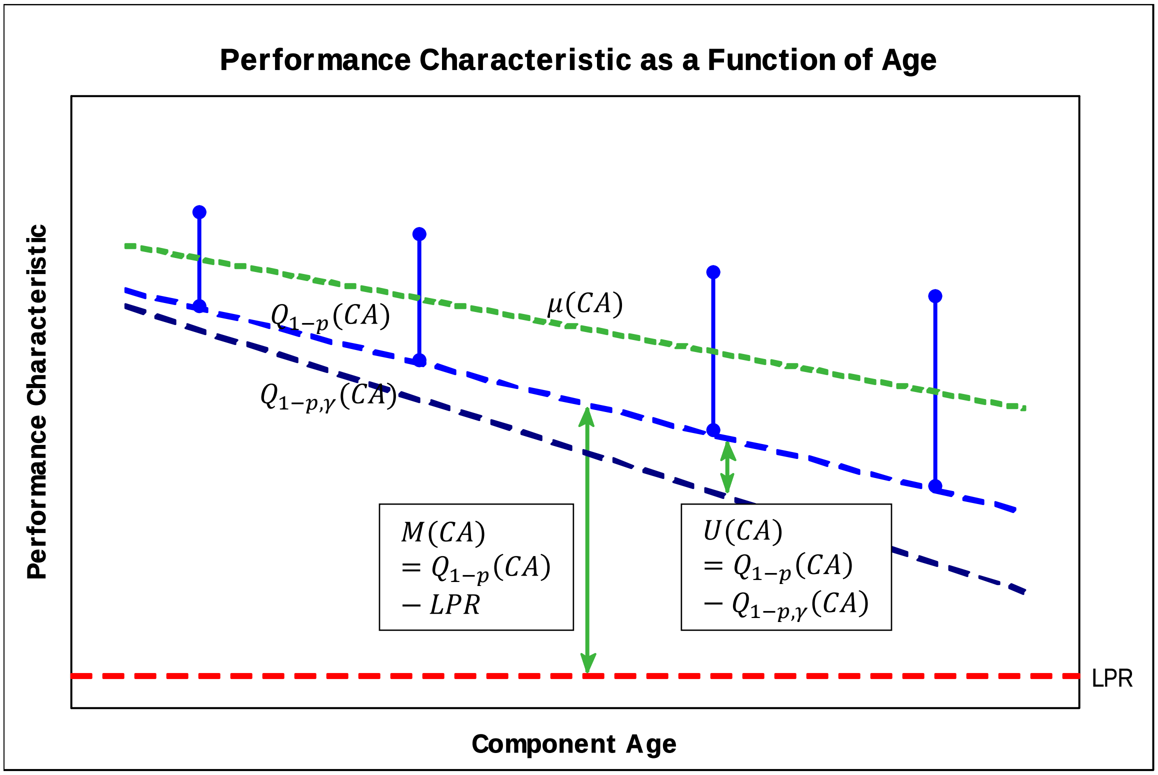

Figure 4.8. Graphical Depiction of Margin and Uncertainty with a Linear Aging Trend.

Again, for simplicity, we will only explore a linear regression model with the component age (CA) as a single independent variable. The regression model is given by,

$$PC = \beta_0 + \beta_1 \cdot CA + \varepsilon,$$

where PC is the performance characteristic, β_0 and β_1 are the model parameters to be estimated, and ε is a random error assumed to follow a Normal distribution with mean zero and standard deviation σ_R. For Example Dataset 2, shown again in Figure 4.9, the model parameters are estimated to be $\hat{\beta}_0 = 10.493$ and $\hat{\beta}_1 = -0.031$ which gives the estimated regression line for the mean performance characteristic at age CA, shown by the dashed green line, to be $\widehat{PC} = 10.493 - 0.031 \cdot CA$. Example Dataset 2 also yields an estimate of $\hat{\sigma}_R = 0.246$.

For a regression analysis assuming the random error follows a Normal distribution with mean zero and standard deviation σ_R, the $r \cdot 100^{\text{th}}$ percentile is a function of the component age and is defined to be,

$$Q_r(CA) = \beta_0 + \beta_1 \cdot CA + \sigma_R \cdot \Phi^{-1}(r),$$

where $\Phi(\cdot)$ is the cumulative probability function for a standard Normal distribution ($\mu = 0, \sigma = 1$). The best estimate of this percentile is obtained by replacing the model parameters with their respective best estimates. Hence,

$$\hat{Q}_r(CA) = \hat{\beta}_0 + \hat{\beta}_1 \cdot CA + \hat{\sigma}_R \cdot \Phi^{-1}(r).$$

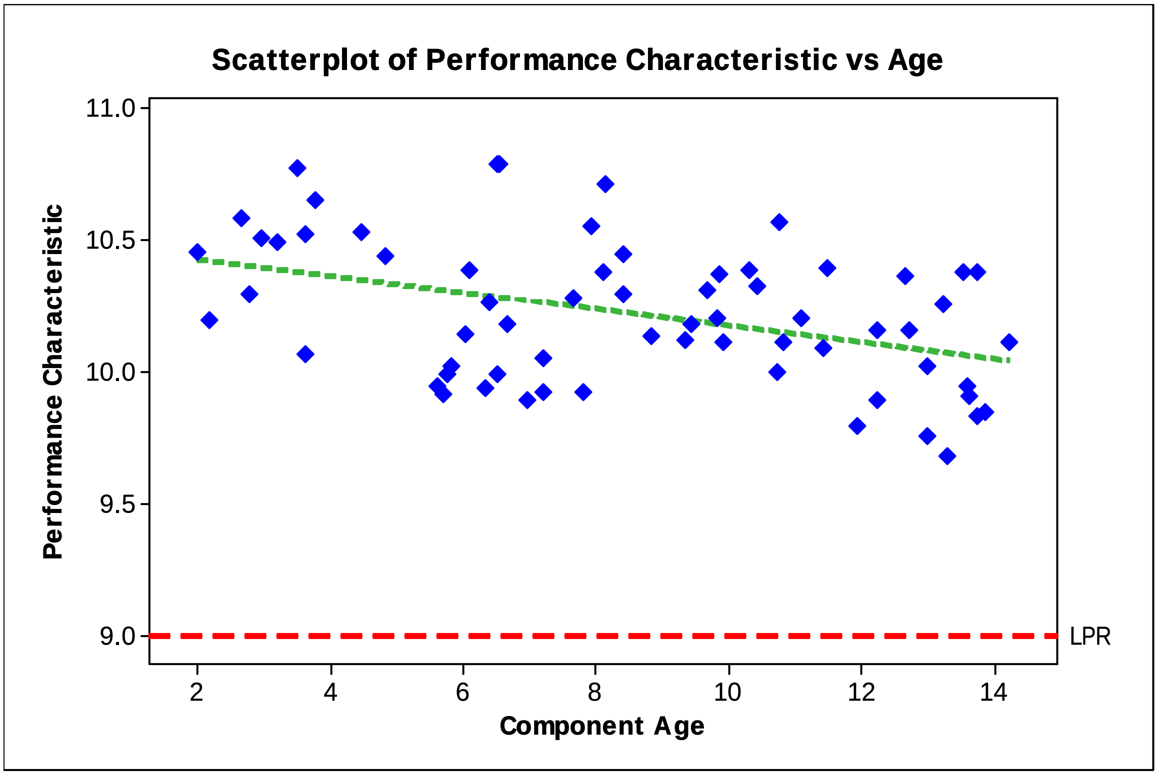

Figure 4.9. Scatter Plot of Example Dataset 2 with Regression Fit.

To be consistent with the example used previously we will again assume we require $P < P_{req} = 0.005$. Since this analysis is specific to a lower performance requirement, the desired percentile would be $\hat{Q}_{1-p}(CA)$. Recall, the content is $p = 1 - P_{req} = 0.995$, hence we are interested in the 0.5th percentile $(1 - p = 0.005)$ and $\Phi^{-1}(0.005) = -2.576$. Therefore, the estimated 0.5th percentile as a function of component age for this set of data is,

$$\hat{Q}_{1-p}(CA) = \hat{Q}_{0.005}(CA) = 10.493 - 0.031 \cdot CA + 0.246 \cdot \Phi^{-1}(0.005) = 9.859 - 0.031 \cdot CA.$$

As discussed above, a confidence bound on a percentile is a statistical tolerance bound. For a regression with a random error that follows a Normal distribution, the estimated lower tolerance bound from a sample of size n is of the form,

$$\hat{Q}_{1-p,\gamma}(CA) = \hat{\beta}_0 + \hat{\beta}_1 \cdot CA - \hat{\sigma}_R \cdot k_1(CA),$$

where

$$k_1(CA) = d(CA) \cdot t_{n-2,\gamma}\left(\frac{\Phi^{-1}(p)}{d(CA)}\right), \quad d^2(CA) = \frac{1}{n} + \frac{(CA - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

x_i is the ith component age, $\bar{x}$ is the average component age, and $t_{df,\gamma}(\Delta)$ denotes the $\gamma \cdot 100^{\text{th}}$ percentile of a non-central t-distribution with df degrees of freedom and noncentrality parameter Δ. The value k_1 is commonly referred to as a tolerance factor. For details on the derivation of this tolerance factor and the tolerance bound in general, please refer to reference 5. For an upper tolerance bound, the appropriate percentile is $\hat{Q}_p(CA)$ and the upper tolerance bound is, $\hat{Q}_{p,\gamma}(CA) = \hat{\beta}_0 + \hat{\beta}_1 \cdot CA + \hat{\sigma}_R \cdot k_1$.

For Example Dataset 2, the average component age is $\bar{x} = 8.56$ and $\sum_{i=1}^{n}(x_i - \bar{x})^2 = (n-1) \cdot s_{age}^2 = 817.856$, where s_{age} is the estimated standard deviation of the component ages. Therefore,

$$d^2(CA) = \frac{1}{65} + \frac{(CA - 8.56)^2}{817.856},$$

and $k_1(CA)$ and $\hat{Q}_{1-p,\gamma}(CA)$ are computed using the equations above. The estimated margin for this dataset is $\widehat{M}(CA) = \widehat{Q}_{1-p}(CA) - LPR = 0.859 - 0.031 \cdot CA$. The estimated uncertainty is $\widehat{U}(CA) = \hat{Q}_{1-p}(CA) - \hat{Q}_{1-p,\gamma}(CA) = \hat{\sigma}_R \cdot (\Phi^{-1}(p) - k_1(CA))$. These metrics are shown in Figure 4.10 along with the fitted regression line. The estimates are projected to the age at which they intersect the lower performance requirement.

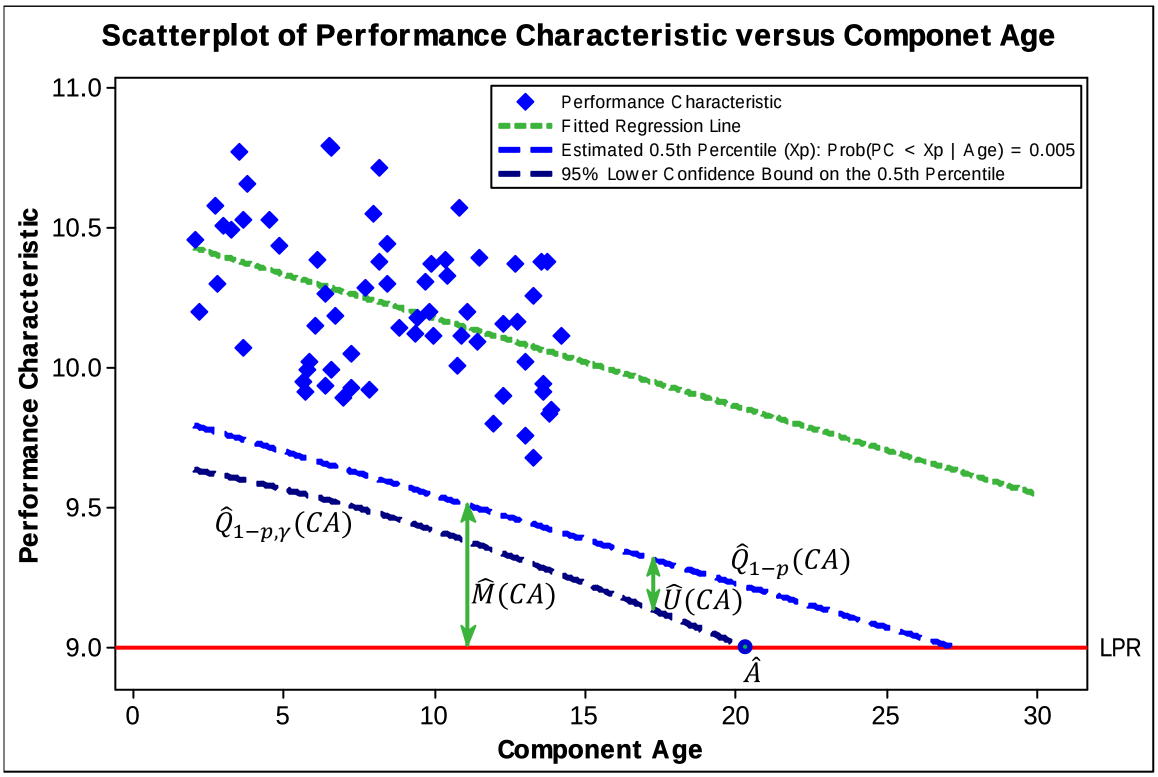

Figure 4.10. Plot of Example Dataset 2 with the Proposed QMU Regression Metrics.

Finally, the age at which the statistical tolerance interval crosses the lower performance requirement, $\hat{A}$, corresponds to the component age at which $\widehat{M} = \widehat{U}$ and $TR = 1$. Therefore, for ages less than $\hat{A}$ we are 95% confident that at least 99.5% of the performance characteristic values will be greater than the lower performance requirement of 9. For ages greater than $\hat{A}$ we cannot make this statement. We define this age to be the component ***Alarm Age***. For Example Dataset 2, $\hat{A} = 20.4$. Recall, the traditional k-factor regression analysis produced a 95% lower confidence bound for the alarm age of $A_{LB} = 21.3$. This new methodology produced an estimate that is almost a year earlier. This is to be expected since the traditional k-factor regression only accounts for the uncertainty in the estimation of the mean, hence the new methodology will always provide a more conservative estimate that accounts for the sampling uncertainty in both the mean and standard deviation.

4.5. Exploration of the Tolerance Ratio

This section is intended to provide further insight into the new tolerance ratio figure-of-merit. In particular, we attempt to provide additional detail on the factors that may cause the tolerance ratio to take on smaller or larger values and an interpretation of how a smaller or larger values of the tolerance ratio influence decision-making. The discussions that follow will be with respect to a lower performance requirement for reference. Similar derivations and statements could be made with minor adjustments for an upper requirement. Further, we only consider the Normal distribution case in the discussion below. Similar comparisons can be made for other distributions but are omitted here for brevity. Recall, the tolerance ratio is defined as,

$$TR = \frac{\widehat{M}}{\widehat{U}} = \frac{\hat{Q}_{1-p} - LPR}{\hat{Q}_{1-p} - \hat{Q}_{1-p,\gamma}} = \frac{\bar{x} - s \cdot z_p - LPR}{\bar{x} - s \cdot z_p - \left(\bar{x} - s \cdot k_1(p,\gamma,n)\right)} = \frac{\bar{x} - s \cdot z_p - LPR}{s \cdot \left(k_1(p,\gamma,n) - z_p\right)},$$

where $z_p = \Phi^{-1}(p)$ is the p^{th} percentile of a standard Normal distribution. It is clear that the tolerance ratio is a function of the content p and the confidence level γ. It should be apparent that as the content increases (a higher proportion of the units required to be within the performance requirements) or the statistical confidence increases (a higher level of rigor required for decision-making) the tolerance ratio will decrease and vice-versa. These parameters however are specific to the performance characteristic and the requirements. For any given analysis these will be fixed, and therefore will be treated as constants in the discussion that follows. For consistency with the previous examples we will fix these at $p = 0.995$ and $\gamma = 0.95$. With these considered constant, the tolerance ratio becomes a function of the mean, standard deviation, and the sample size,

$$TR = \frac{\bar{x} - s \cdot z_p - LPR}{s \cdot \left(k_1(n) - z_p\right)}.$$

For a fixed content p and confidence level γ, $k_1(n)$ is proportional to $1/\sqrt{n}$, denoted by $k_1(n) \propto 1/\sqrt{n}$. Therefore, if we consider the tolerance ratio to be a function of any one of these three parameters (the mean, standard deviation, and the sample size) holding the other two fixed, we have $TR(\bar{x}) \propto \bar{x}$, $TR(s) \propto 1/s$, and $TR(n) \propto \sqrt{n}$. These three relationships are shown in Figure 4.11 below.

Using these relationships, we can explore how to increase the tolerance ratio. The list shown here discusses possible practical ways to increase the tolerance ratio by improving either the component's performance or the amount of information obtained on the component.

1. Increase margin – Improvements in the component performance (possibly through design improvement) that shift the mean of the distribution of the performance characteristic away from the performance requirement will result in a larger tolerance ratio, **provided the unit-to-unit variability does not change significantly**.

2. Decrease uncertainty – Reducing the unit-to-unit variability of the component performance distribution (possibly through improvement or implementation of manufacturing process controls) will result in a larger tolerance ratio, **provided these changes do not shift the mean of the distribution significantly**.

3. Increase precision - Increasing the data acquired on a performance requirement (possibly through increased surveillance testing) will result in a larger tolerance ratio, **provided the additional data does not significantly change the estimated mean or standard deviation**.

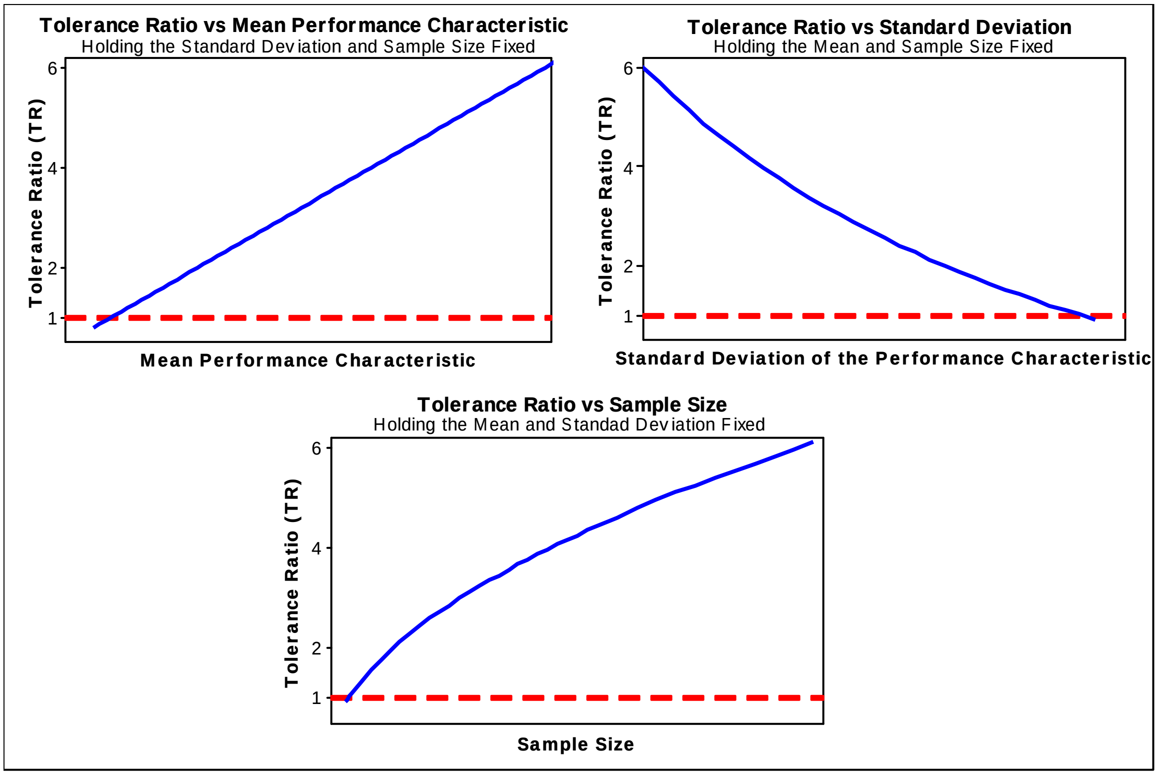

Figure 4.11. The Tolerance Ratio as a Function of the QMU Metrics.
Mean (top left), Standard Deviation (top right), and Sample Size (bottom)

These relationships should be fairly obvious since the tolerance ratio is a ratio of the margin and uncertainty. Recall, the assessed performance is defined as the estimated percentile of the performance characteristic distribution corresponding to a maximum probability of failure, $\hat{Q}_{1-p}$ or $\hat{Q}_p$ for either a lower or upper requirement respectively. The margin is a function of the assessed performance, which is directly related to both the mean and standard deviation. As the mean of the performance characteristic distribution improves relative to the performance requirement, the assessed performance, and hence the tolerance ratio, will increase. Likewise, the assessed performance and the tolerance ratio increase as the standard deviation decreases. In either case the tolerance ratio increases as the assessed performance increases, however this is not a one-to-one relationship and will depend on which parameter (the mean or standard deviation) is driving the change in the assessed performance. Figure 4.12 shows these relationships for changes in the mean and standard deviation. These are presented because it may be more meaningful to think about changes in the tolerance ratio as a function of the assessed performance rather than simply as a function of either the mean or standard deviation alone.

Note that no plot is shown in Figure 4.12 for a change in the sample size since the assessed performance is not directly a function of the sample size. Recall number 3 above noted that increasing the data acquired on a performance requirement would result in a larger tolerance ratio, **provided the additional data does not significantly change the estimated mean or standard deviation**. The change in the tolerance ratio in this case is due to the fact that the **precision** of the assessed performance (the estimated percentile) will get better (i.e. $\hat{U}$ is smaller). **There is however no guarantee that the assessed**

performance will remain the same. The estimated percentile may increase or decrease as new data is acquired. If the estimated percentile moves closer to the performance requirement (gets worse) then the tolerance ratio may go down even as the precision of the estimate gets better. This would be the case where, although the uncertainty gets better (smaller), the margin gets worse (smaller). This relationship is difficult to show graphically, however this discussion is included as a warning that simply increasing the sample size may not necessarily improve or change the resulting conclusions.

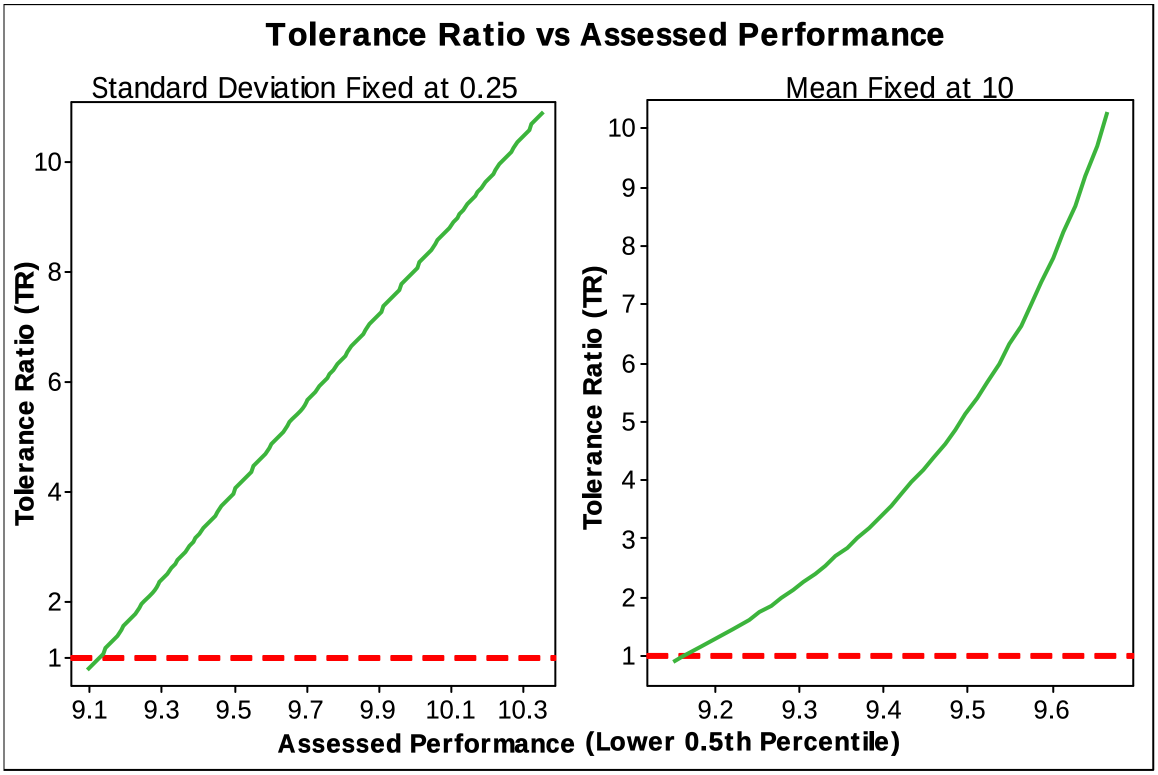

Figure 4.12. The Tolerance Ratio as a Function of the Assessed Performance.
As the Mean Changes (left) and as the Standard Deviation Changes (right)

The plots shown to this point in this section provide some insight as to the factors that influence changes in the tolerance ratio, however they do not provide much insight as to what the magnitude of the tolerance ratio implies with respect to performance. This insight is ultimately required to make comparisons across components. For simplicity, we only consider comparisons of the tolerance ratio for a fixed content p and confidence level γ. Recall the concept of a coverage probability introduced in Section 4.1. The coverage probability is defined as the estimated probability that a performance characteristic will be either greater than a lower requirement or less than an upper requirement with $\gamma \cdot 100\%$ confidence. That is, the coverage probability is the value p_c that satisfies the equation, $\hat{Q}_{1-p_c,\gamma} = LPR$ for a lower requirement or $\hat{Q}_{p_c,\gamma} = UPR$ for an upper requirement. In other words, it is the content value such that the tolerance bound will be exactly equal to the performance requirement.

First, we make the following observation, again with the content and confidence level fixed at $p = 0.995$ and $\gamma = 0.95$. When the tolerance ratio is equal to one, the estimated tolerance bound will be equal to the performance requirement and the coverage probability will be equal to the content. If the tolerance ratio is less than one, then the coverage probability will be less than the content and similarly if the tolerance ratio is greater than one the coverage probability will be greater than the content. That is,

$$\begin{array}{lcl} TR < 1 & & p_c < p = 0.995 \\ TR = 1 & \Leftrightarrow & p_c = p = 0.995 \\ TR > 1 & & p_c > p = 0.995 \end{array}$$

We note that the coverage probability is a function of both the assessed performance and the sample size. The following two figures show the coverage probability as a function of the tolerance ratio holding one of these two factors fixed. The curves differ slightly, however they both provide the same insight. As the tolerance ratio increases so does the coverage probability.

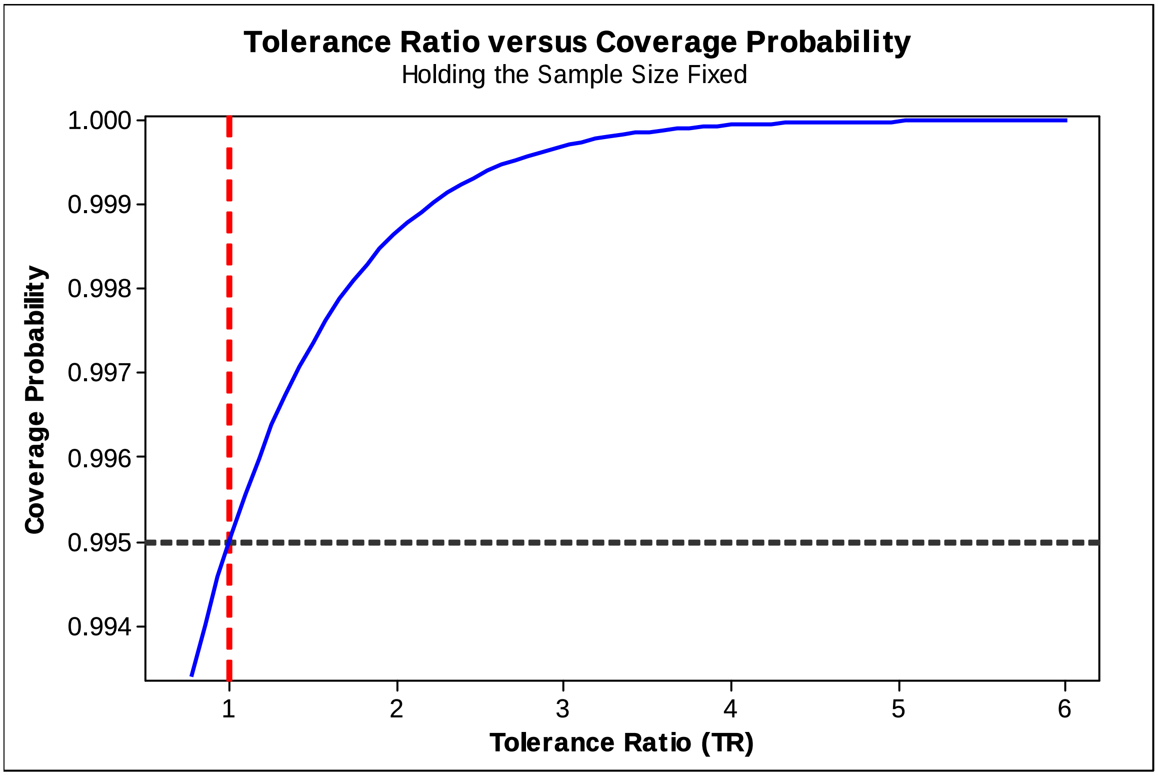

Figure 4.13. The Coverage Probability as a Function of the Tolerance Ratio Holding the Sample Size Fixed.

It should be clear from Figure 4.13 that as the tolerance ratio goes to infinity (which for a fixed sample size would require the assessed performance to go to infinity) the coverage probability goes to one. This is the case where the performance characteristic distribution moves significantly far from the performance requirement, such that the probability that a value would fall outside of the performance requirement becomes negligible. One thing to note from Figure 4.14 is that for a fixed assessed performance, an increase in the sample size will provide more confidence in the estimated parameters and hence the tolerance ratio and the coverage probability will both increase. The coverage probability however will not approach one as the tolerance ratio approaches infinity as it will in Figure 4.13. The coverage probability in this case is limited by the true performance. In the example shown here, the mean and standard deviation are fixed at $\bar{x} = 10$ and $s = 0.25$ which, if we assume these parameters to be known (which is the case if $n \to \infty$) gives an estimated probability that the performance characteristic will be greater than the lower performance requirement of 9,

$$Prob(PC > 9) = 1 - \Phi\left(\frac{9 - 10}{0.25}\right) = 0.99997.$$

Therefore, if only the sample size changes, the coverage probability will only approach 0.99997 as the tolerance ratio goes to infinity (again assuming that the mean and standard deviation do not change).

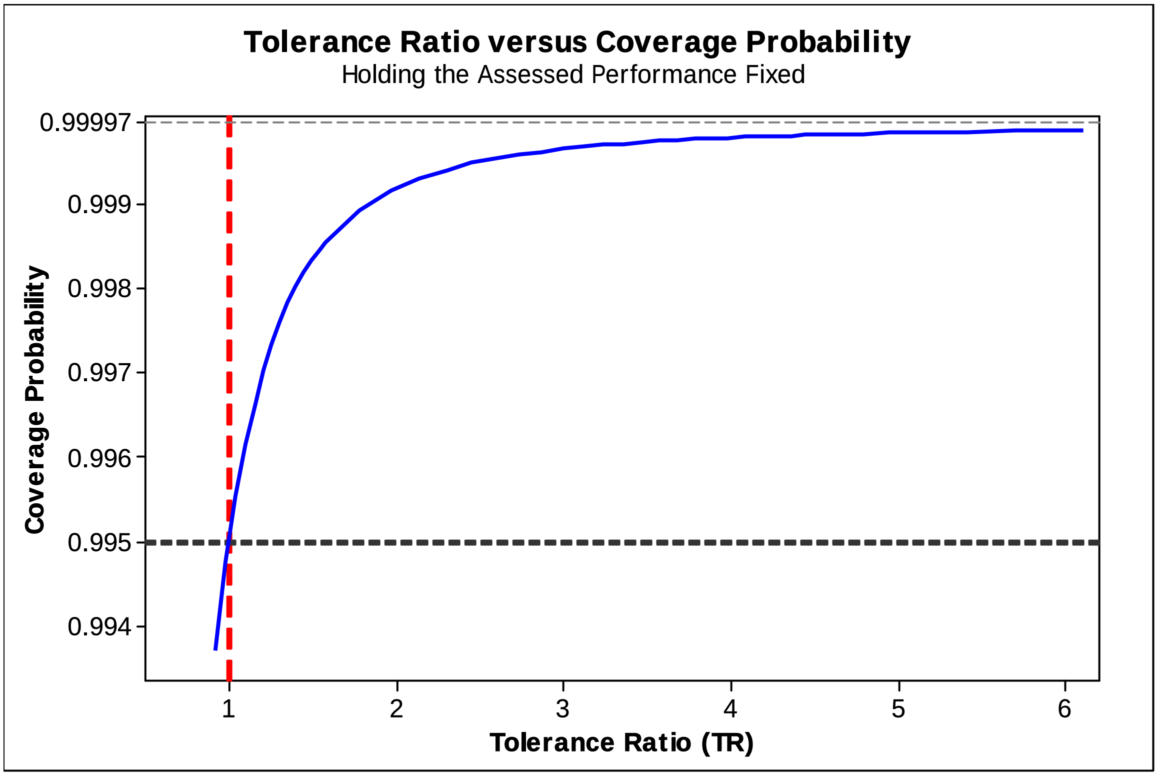

Figure 4.14. The Coverage Probability as a Function of the Tolerance Ratio with the Assessed Performance Fixed.

Figure 4.13 and Figure 4.14 should make it clear that as the tolerance ratio increases so does the coverage probability. Therefore, if we compare two components, each with the same maximum probability of failure requirement and confidence level, the one with the larger tolerance ratio will also have a larger coverage probability; hence that component will have a smaller probability of failing to meet the performance requirement. The computation of the coverage probability will aid in these types of comparisons, especially for analyses resulting in a tolerance ratio close to one. For analyses resulting in a large tolerance ratio ($TR \gg 5$), small changes in a tolerance ratio of this magnitude will have negligible differences in the coverage probability. For examples of the computation of the coverage probability refer to Sections 4.3.1 and 4.3.2.

5. PROPOSED ACTIVITIES TO BE INCLUDED IN AN ENGINEERING DATA ANALYSIS

This section describes a thorough engineering analysis of a physical dataset. The intent of an engineering analysis is to better understand the available data (and its nuances) and to identify a suitable analysis set for further investigation and analysis. The engineering analysis should be performed well in advance of any planned statistical analysis since this step will often require several iterations and discussions with systems and component engineers, tester engineers, analysts, and statisticians. Including newly acquired data after this process has begun is not recommended unless the new data is subjected to an engineering analysis consistent with the initial engineering analysis. Simply adding in newly acquired data to an analysis after the engineering analysis has been completed is risky since the new data may not be understood at the same level of rigor as the data that went through the engineering analysis.

This section is also intended to reintroduce the original intent of QMU and identify steps that will help component engineers and analysts proceed in the right direction. Recall, the three key elements of QMU consist of the following: (1) specification of performance thresholds; (2) identification of associated performance margins, where a performance margin is a measure of exceeding the performance threshold; and (3) quantification of uncertainty in the performance thresholds and performance margins as well as in the larger framework of the decisions being contemplated. Diegert et al. [1] defines these three key elements, condensed below, as follows.

- **Performance threshold:** *Performance* is the ability of a bomb, a warhead, or a component to provide the proper function (e.g., timing, output, response) when exposed to the sequence of design environments and inputs. This definition of performance is applicable to the following functional-requirement areas: reliability, nuclear safety, use-control, and nuclear survivability. A *performance threshold* is a specification of a necessary performance achievement, typically in quantitative form. We call this the *required performance* of a system. The required performance is most often specified in a deterministic form where the performance must be greater than (or less than) a specified performance threshold.

- **Performance margin:** A performance margin is the difference between the required performance of a system and the demonstrated performance of a system, with a positive margin indicating that the expected performance exceeds the required performance. The expected performance can be nondeterministic and may be specified by a probability or cumulative distribution function.

- **Uncertainty:** There is uncertainty in the specification of thresholds and margins, as well as in the larger framework of the decision tasks. This uncertainty begins in the requirements that provide a foundation for the definition of performance thresholds, and it accumulates and transforms as the various science and engineering activities that lead from weapons design to qualification to evaluation are executed. There are two general types of uncertainty that must be separately accounted for, quantified, and aggregated within QMU:

    1. *Aleatory uncertainty* – also called irreducible uncertainty or stochastic variability. We typically refer to this type of uncertainty as simply *variability*. Aleatory uncertainty (or variability) is naturally characterized, quantified, and communicated in terms of probability.

    2. *Epistemic uncertainty* – also called reducible uncertainty. This type of uncertainty is due to lack of knowledge or incomplete knowledge. Common examples of epistemic uncertainty are the so-called model form uncertainty (that is, uncertainty in how well the equations in the

model capture the physical phenomena of interest), both known and unknown unknowns in scenarios, and poor-quality physical test data.

With these key elements serving as the basis for a rigorous QMU analysis, the following steps are recommended for an engineering analysis. The steps are presented in the order in which the engineering analysis is proposed to occur, however it may be necessary to revisit some steps based on the decisions made at a later step.

1. Identify the performance characteristics that relate to or provide knowledge of system and/or component performance. Time should be spent at this step to understand how the performance characteristics are measured and the testing conditions used to obtain the measurements. This step could also include a failure modes and effects analysis (FMEA), or a similar type of analysis, to identify if the current data is sufficient for assessing component performance. The testing conditions and measurements should be an accurate representation of component function in use conditions. All measured values that do not relate to performance and/or are not obtained in a well-understood fashion that can be related to component performance and thresholds should be excluded from more rigorous QMU analyses. The measurement uncertainty (especially the stability of the measurements over time when trending is being considered) should be investigated at this step as well.

2. For each performance characteristic remaining after step one, time should be spent understanding the performance thresholds and requirements and the planned lifetime of the component. All characteristics that do not have a well-defined and meaningful performance requirement, which relates directly to component and/or system performance, should be removed from further rigorous QMU analyses until the requirements and thresholds are better understood. Further, understanding the planned lifetime of the component is necessary for all age trend analyses in order to provide insight as to the range of age values that the age trend must be projected across.

3. Investigate the composition of the available data. The units tested should be a **representative** and **random** sample from the stockpile population that inferences are intended to be made on. If convenience sampling (sampling based on availability/cost of resources) is used one must be aware of the potential limitations of that sample. If the sample only represents a subset of the overall population then the results and conclusions should only be applied to that subset. If the pedigree of all or a subset of the data is in question, then more rigorous QMU analyses should not be performed until the dataset is properly understood or reduced to an appropriate analysis set. If pedigree differences are observed within the dataset work should be done to understand and identify a subpopulation that best represents the stockpile population, or representing a worst-case subpopulation. Ideally, further more rigorous analyses should only be performed on the most representative dataset.

4. Perform a preliminary graphical analysis on the remaining performance characteristics to identify potential data issues such as shifts or jumps in the data by production date, test date, test code, or other potential factors. If any of these factors are identified, work should be done to better understand the impact of the observed factor(s). Simply dividing the data into several subsets based on observed shifts is not recommended. Recall, the aleatory uncertainty accounts for variability arising from a number of effects (unit-to-unit differences, lot-to-lot differences, lack of measurement precision, etc.), none of which should be dominant. If an underlying factor dominates the aleatory uncertainty, the uncertainty must be properly characterized in order for the conclusions from subsequent analyses to be meaningful. The decision as to how to handle these additional factors should be based on engineering judgment. If a shift is caused by a change to a tester then this difference needs to be understood and accounted for prior to further analysis. If a

new tester is assumed to be a better representation of use conditions and the cause of the observed shift from the old tester is unknown then only data from the new tester should be considered for analysis. If however, the observed shifts are due to test code differences, possibly indicating sensitivity to environmental factors, one may wish to either analyze the environmental conditions separately or account for them in a more rigorous statistical analysis such as an analysis of covariance. Consultation with a statistician is highly recommended at this step. Further rigorous QMU analyses should not be performed until all factors have been thoroughly investigated.

5. During the preliminary graphical analysis, identify any performance characteristics with insufficient data to perform a more rigorous analysis on. Tests with high variability, a large number of factors, and a small sample size (e.g. flight tests) are not appropriate for a rigorous statistical analysis. These tests can provide valuable engineering information and could potentially aid a computational simulation QMU analysis, however the data alone should not be analyzed using the methods described in this paper for a physical simulation QMU analysis.

6. Identify performance characteristics with ample margin. This often can be achieved by generating a scatter plot of the data along with the performance requirements. If this plot shows that both the margin and uncertainty (variability) of the observed data are sufficient as compared to the requirements, further more rigorous analyses are not needed. The traditional k-factor based on the mean and standard deviation and its lower confidence bound can also be used as a guide to determine if the observed margin warrants further analysis. One could potentially set a cutoff to exclude performance characteristics with large margin from further analysis. For example, any performance characteristic with a lower confidence bound on the k-factor greater than five $\left(\hat{k}_{\gamma} > 5\right)$. This cutoff however should be used cautiously since other factors, especially the distributional assumption, can affect the interpretation of the k-factor.

7. Once the dataset has been reduced to a set of performance characteristics that are suitable for further analysis based on the steps above, the remainder of the engineering analysis should focus on verifying assumptions, identifying potential outliers and deciding how to handle them, and identifying potential aging trends. This step can often be completed as part of the preliminary graphical analysis. Here potential aging trends should be investigated to see if other underlying factors might be contributing to the trending. All potential outliers should be taken seriously and should be investigated thoroughly for explanations. Automatic outlier rejection schemes such as throwing out all data beyond 4 sample standard deviations from the sample mean are particularly dangerous and not recommended. Determining if outliers should be removed from the analysis set is a process that should include interaction between the statistician/analyst, the component engineers, and the tester group.

The steps presented above are a good starting point for a thorough engineering analysis however; other steps may be needed as the study of the data and its nuances progresses. In particular, the completion of one step may prompt the analysis to revisit a previous step. Once the engineering analysis is complete, the component engineer and analyst should thoroughly understand the data to be analyzed further. Again, upon completing the engineering analysis of the data, an ***Engineering Review*** must be performed by an interdisciplinary team (Systems and Components Engineering, Test Engineers, Statisticians and Analysts, etc.) to ensure that the conclusions resulting from the engineering analysis are acceptable and the dataset is sufficient for further analysis. Although only a small subset of the data and performance characteristics may be appropriate for further analysis, a good deal of information can be obtained from the engineering analysis and review itself. This study can indicate areas with knowledge gaps that need further investigation or performance characteristics with insufficient data, and lead to decisions on future research and/or a formal design of experiments and collection of new data to fill existing data gaps.

6. A COMPARISON TO QMU FOR COMPUTATIONAL SIMULATION APPLICATIONS

This new approach to QMU for physical simulation data is closely aligned to the approach to QMU for computational simulation analyses. This consistency will simplify analyses involving both physical and computational simulation and will help clarify Sandia's approach to QMU. The approaches are identical in their objectives but differ somewhat in the methodology employed because of the, often substantial, contribution of epistemic uncertainty that is common with computational analyses. In comparing the two approaches, we assume a calibrated computational model that has successfully completed application-specific verification and validation analysis and testing. **This is equivalent to the assumption that an engineering analysis has been performed on a dataset prior to applying the methodologies proposed for a QMU on physical simulation data.** While these difficult elements of establishing reliable model-based predictions are an integral part of computational simulation QMU, we focus here on the comparisons of approaches to establishing margins and uncertainty. Furthermore, for simplicity in this comparison, we will only address the point-in-time analysis and address only the computational simulation case where there is a probabilistic characterization for the epistemic uncertainty. Other situations have been addressed [7,8] but are beyond the scope of this document.

The computational simulation program is currently addressing QMU using an approach known as ***Probability-of-Frequency*** (previously referred to as "second-order") method. Technical details concerning this approach are given in reference 7. In this approach, the simulations that provide the basis for prediction are selected through a two-stage sampling approach as illustrated in Figure 6.1. In the inner loop, the simulations are performed using fixed values for the epistemic parameters, producing a single distribution of the performance characteristic conditioned on the epistemic values.

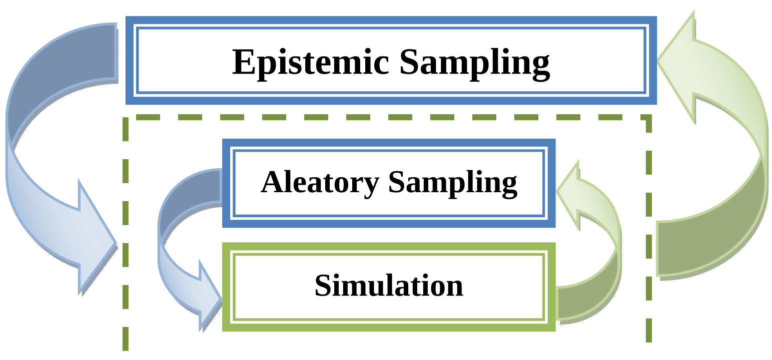

Figure 6.1. Looping Process for the Probability-of-Frequency Approach.

One replication of this inner loop will generate data for a performance characteristic yielding a histogram similar to that shown in Figure 6.2. Often, for computational analyses, this empirical distribution is used directly (as opposed to trying to fit a parametric distribution) although the latter has been proposed for cases where there are few (expensive) simulations. We assume here (to simplify this comparison) that the simulations are inexpensive. Alternative techniques for implementing the probability-of-frequency approach with fewer simulation runs are currently being examined [9]. The fitted parametric distribution shown in Figure 6.2 and Figure 6.3 is shown for reference; however the calculations that follow will use the empirical distribution directly.

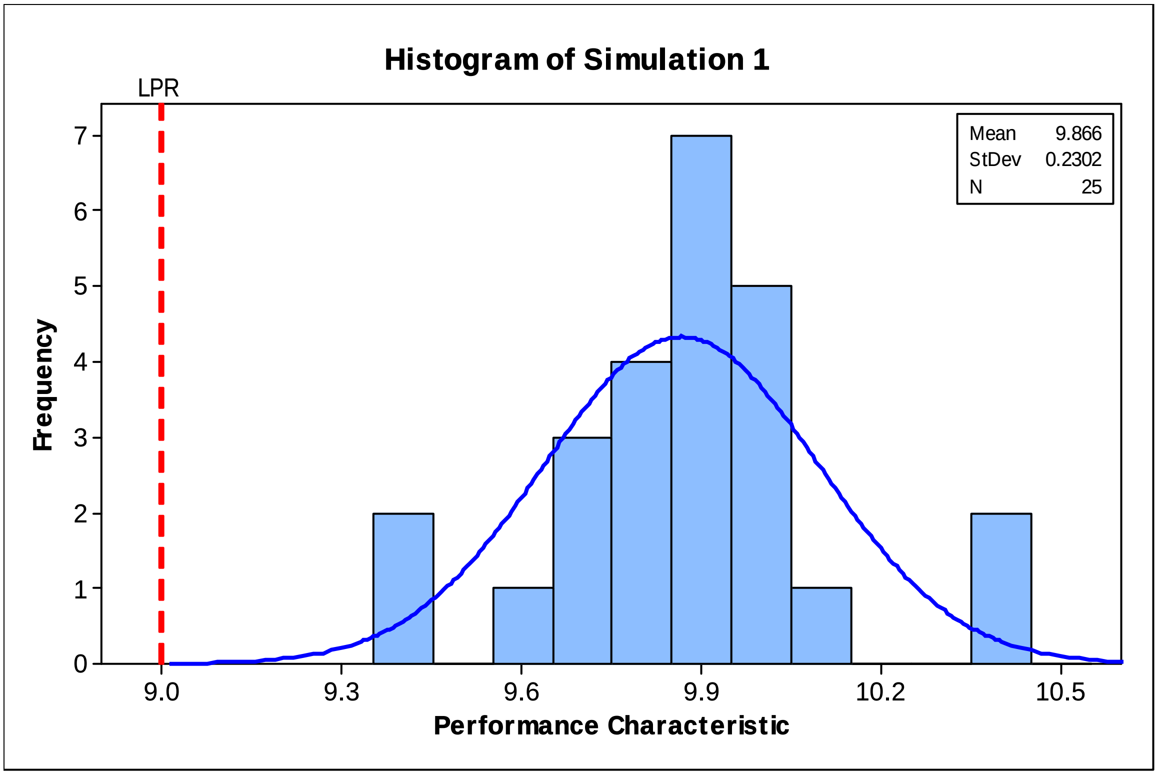

Figure 6.2. Histogram of a Single Simulation from the Inner Loop.

The data shown in Figure 6.2 is similar to the physical simulation data (Example Dataset 1) shown in Figure 2.3 except that the data was acquired through a single inner loop simulation of a physics-based computer model. It is assumed that this inner loop properly characterizes the aleatory uncertainty. In the absence of additional epistemic uncertainties, this single set of simulation data could be subjected to the same statistical methodologies proposed in Section 4 to estimate margin and uncertainty. This however treats the computational simulation data exactly as physical simulation test data and does not necessarily account for inaccuracies, often present, in the implemented physics model.

Since multiple simulations will be performed for various instances of the epistemic parameters, it is often convenient to represent the data graphically using an empirical cumulative distribution function (CDF). Figure 6.3 shows an empirical CDF of the data from Figure 6.2. The inner loop is repeated several times for various sampled epistemic values in the outer loop. Figure 6.4 shows a series of simulations, represented by their empirical CDF's, collected for the given epistemic parameters sampled in the outer loop. There are different methods for selecting these values but random sampling is assumed here. Figure 6.4 also shows a lower performance requirement (red dashed line) and the maximum proportion of units allowed to fall below the lower performance requirement (black dotted line). These are equivalent to LPR and P_{req} respectively from the discussions above for physical simulation data. The red dots indicate the estimated percentile of each empirical distribution corresponding to P_{req}. The objective in the probability-of-frequency approach is to estimate the distribution of the desired percentile. In the physical simulation case, this distribution is generally calculated based on an assumed parametric distribution fit to the data from a number of physical experiments and is quantified by the calculated statistical tolerance bound. For consistency in comparing back to the previous examples, the example shown in Figure 6.4 estimates the 0.5^{th} percentile for each inner loop simulation.

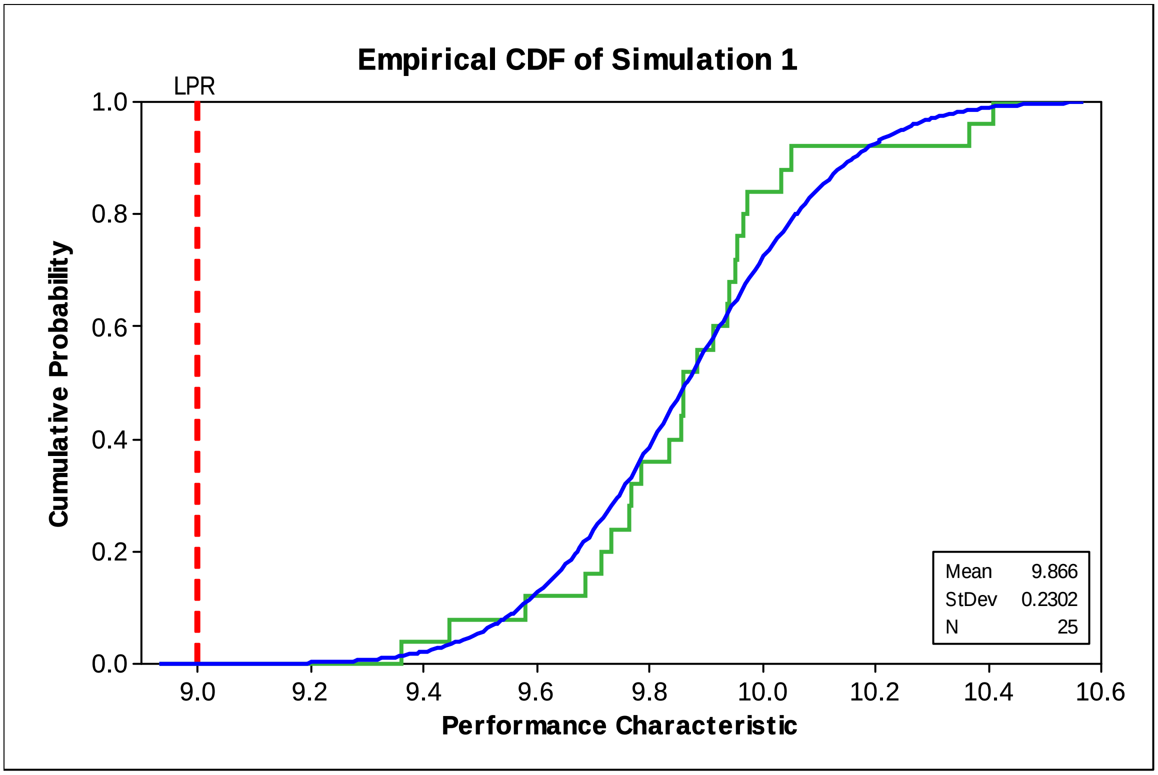

Figure 6.3. Empirical CDF of a Single Simulation from the Inner Loop.

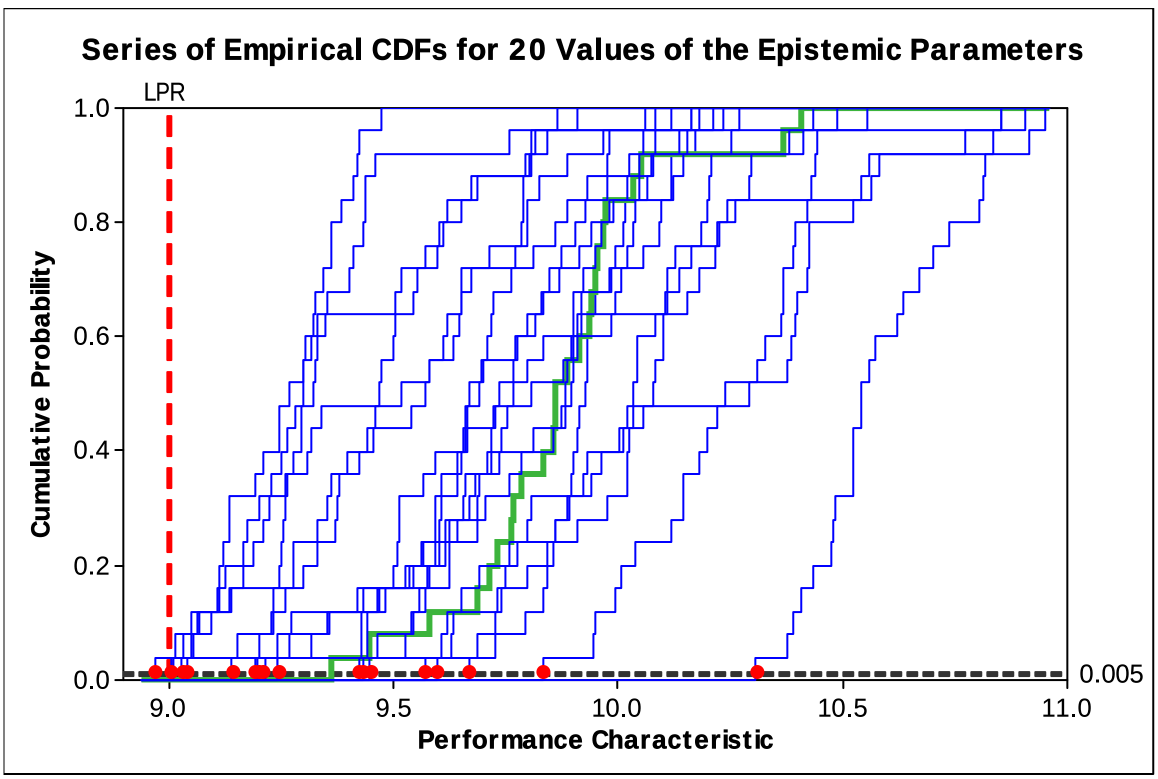

Figure 6.4. Resulting Empirical CDFs Using the Probability-of-Frequency Approach.

The epistemic uncertainty distribution for the 0.5th percentile, shown in Figure 6.5, is constructed empirically using the estimated points indicated by the red dots. Again, the points are estimated to satisfy the probabilistic requirement (the maximum proportion of units allowable below the lower performance requirement of 0.005) conditioned on each set of values of the epistemic parameters.

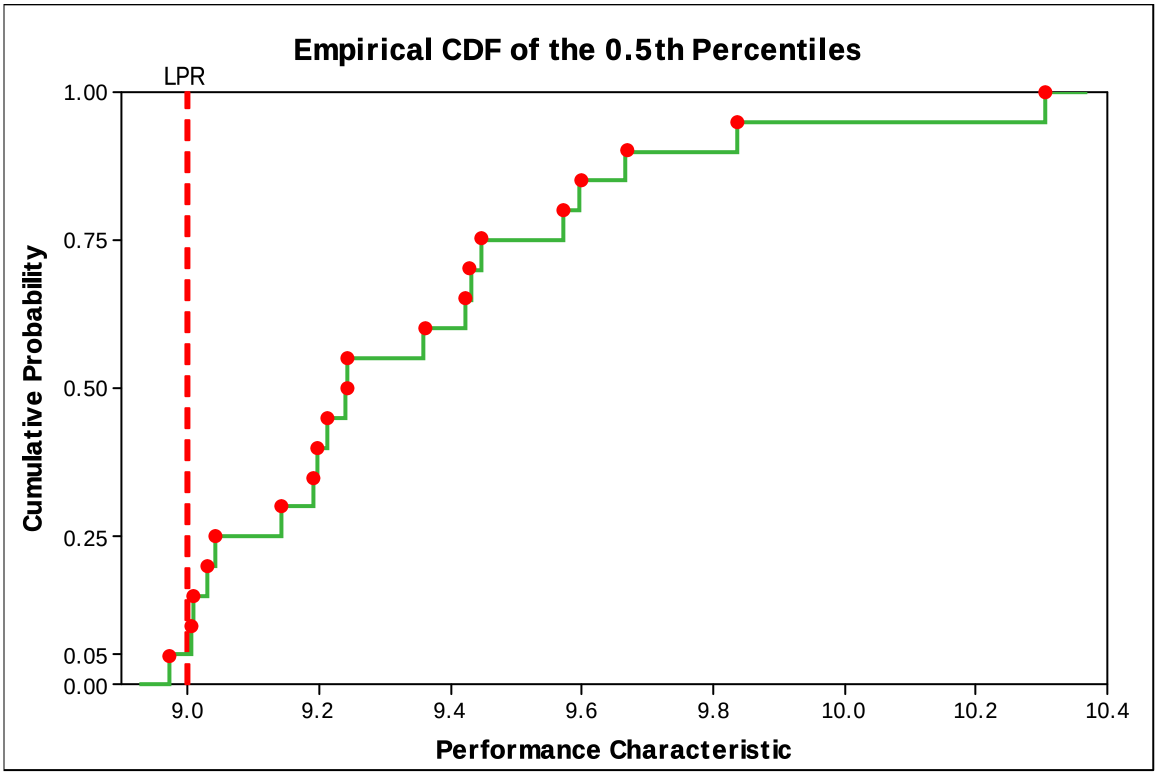

Figure 6.5. Uncertainty CDF Using the Probability-of-Frequency Approach.

Each red dot in Figure 6.4 and Figure 6.5 is an estimated lower 0.5th percentile, $\hat{Q}_{0.005}$, from a single iteration of the inner loop. The goal then becomes the quantification of margin and uncertainty for this distribution relative to the lower performance requirement. This was accomplished in the proposed methodology for physical simulation data by computing an estimate of the percentile and a lower confidence interval for the estimated percentile (a statistical tolerance bound). The probability-of-frequency approach described here produces a distribution of the estimated percentiles. Therefore, the average value from this distribution can be used as a point estimate of the percentile. We denote this by $\bar{Q}_{1-p}$ or $\bar{Q}_p$ for a lower or upper percentile respectively. The *margin* is then defined as the difference between this point estimate and the performance requirement, $\widehat{M} = \bar{Q}_{1-p} - LPR$ of a lower requirement or $\widehat{M} = UPR - \bar{Q}_p$ for an upper requirement. Next, to account for the uncertainty in this point estimate a confidence bound can be obtained from the epistemic uncertainty distribution by selecting the appropriate percentile of the distribution. For consistency with the previous examples, we choose a 95% confidence level ($\gamma = 0.95$). We denote this here by $\bar{Q}_{1-p,\gamma}$ and the *uncertainty* can be quantified by the difference between this bound and the estimated mean of the epistemic uncertainty distribution, $\widehat{U} = \bar{Q}_{1-p} - \bar{Q}_{1-p,\gamma}$, specifically for a lower percentile. Similar metrics could be defined for an upper percentile. Figure 6.6 shows these metrics for the example described above.

For this example, the mean percentile from the epistemic uncertainty distribution is $\bar{Q}_{1-p} = 9.35$. The margin, relative to the lower performance requirement of 9, is then estimated to be $\widehat{M} = \bar{Q}_{1-p} - LPR = 9.35 - 9 = 0.35$. The lower 95% confidence interval from the empirical CDF of the 0.5$^{\text{th}}$ percentile is $\bar{Q}_{0.005,0.95} = 8.97$ and the uncertainty is estimated to be $\widehat{U} = \bar{Q}_{1-p} - \bar{Q}_{1-p,\gamma} = 9.35 - 8.97 = 0.38$. Clearly, $\widehat{M}/\widehat{U} = 0.92$, which is less than 1. In the computational simulation framework M/U is referred to as a confidence factor (CF). For a confidence factor greater than 1 it is assumed that unreliability due to margin insufficiency is negligible (see reference 1). For $CF < 1$, as is the case with this example, we cannot make that claim. This decision rule and interpretation are closely aligned with the proposed approach using the tolerance interval methodology for QMU with physical simulation data. This provides a consistency across methods that allows for comparisons and interpretations to be made more easily.

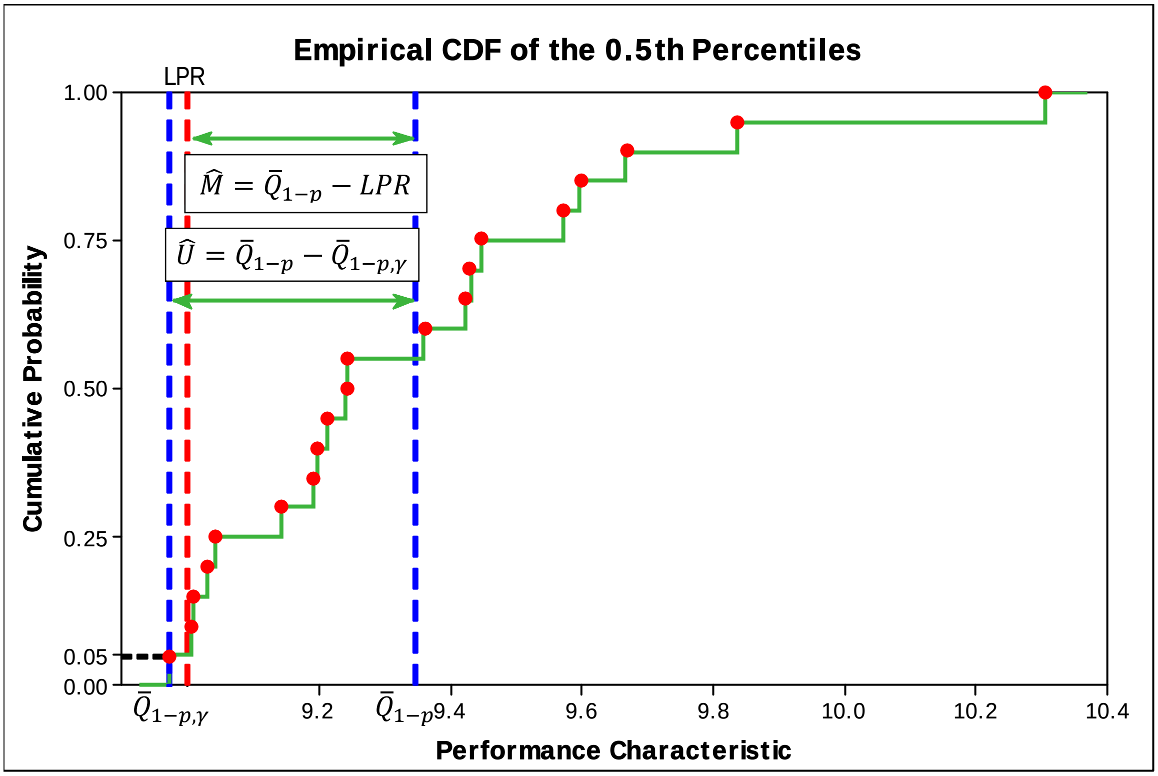

Figure 6.6. Uncertainty CDF Showing the M and U Metrics.

7. CONCLUSIONS

This paper reviewed the standard k-factor methodologies and discussed potential limitations. The tolerance interval methodology was introduced as an alternative technique for QMU with physical simulation data. The technical details of these methodologies were presented and demonstrated with several examples. The proposed methodologies are intended for a thoroughly understood dataset with a performance characteristic that relates to component and/or system function and a well understood performance requirement. This paper also outlined recommendations for an engineering analysis that will result in a dataset that meets these criteria and is eligible for a rigorous QMU analysis using the proposed methodologies. Finally, the tolerance interval methodology was shown to be more consistent with the standard QMU methodologies for computational simulation applications. We conclude with the following recommendations for future QMU analyses on physical simulation data.

1. The point-in-time k-factor analysis should be used as a screening tool in an engineering analysis to identify performance characteristics with low margin that require a more rigorous QMU analysis. This methodology can be applied quickly on a large number of performance characteristics, but should not be viewed as a rigorous analysis technique. This methodology should be used cautiously, especially in cases of non-Normality. Analyses that show a potential for low margin or deviations from Normality should adhere to the guidance in list item 2 below.

2. For performance characteristics that do not exhibit an aging trend but do show the potential for low margin or non-Normality, the point-in-time tolerance interval methodology should be used to assess the impact of the low margin and to make conclusions with respect to requirements. In these cases, considerable thought should go into the choice of statistical distribution and the appropriate methodology should be applied based on this decision. The resulting analysis metrics can easily be presented on the engineering unit scale regardless of the choice of distribution. Consultation with a statistician at this point is highly recommended.

3. For cases where the k-factor regression methodology is currently being implemented, a transition to the tolerance interval methodology is recommended. For analyses with large sample sizes these methods provide comparable results which should make this change straightforward. The tolerance interval approach increases the technical rigor and interpretability of these analyses. The k-factor regression analysis is not recommended moving forward due to a continued lack of understanding and interpretability. A statistician or the QMU steering committee can provide guidance on how to educate customers on this transition.

4. For early screening analyses, where the goal is to assess if one or more of the performance characteristics has an aging trend, standard regression techniques should be used to determine if the trend is statistically significant. A p-value less than 0.05, from a hypothesis test on the slope parameter, is commonly accepted as sufficient evidence of a statistically significant trend. Determination that the trend is a result of an aging effect should be done in an engineering analysis. Cases that show a statistically significant trend, that is determined to be due to an aging effect, should adhere to the guidance in list item 5 below.

5. For performance characteristics that exhibit an aging trend, the tolerance interval regression analysis should be used to estimate an alarm age. This method parallels the methodology for a rigorous point-in-time analysis except for the addition of an aging trend. This provides a more consistent and interpretable approach to QMU for physical simulation data. Again, the resulting analysis metrics can easily be presented on the engineering unit scale with a single graphic.

8. REFERENCES

1. K.V. Diegert, S.E. Klenke, G. Novotny, R.A. Paulsen, M. Pilch, and T.G. Trucano, *Toward a More Rigorous Application of Margins and Uncertainties within the Nuclear Weapons Lifecycle—A Sandia Perspective*, SAND2007-6219, Sandia National Laboratories, Albuquerque, NM, December 2007.

2. J.T. Newcomer, QMU *101 – Just the Basics: The Why's & What's*, PowerPoint Presentation, Sandia National Laboratories, Albuquerque, NM, Last Revised September 2012.

3. J.T. Newcomer, *QMU 102 – Introduction to QMU Statistical Methodologies*, PowerPoint Presentation, Sandia National Laboratories, Albuquerque, NM, Last Revised September 2012.

4. N.L. Johnson, *Systems of Frequency Curves Generated by Methods of Translation*, Biometrika, Volume 36, Number ½, pp. 149-176, June 1949.

5. K. Krishnamoorthy and T. Mathew, *Statistical Tolerance Regions, Theory, Applications, and Computation*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 2009.

6. D.B. Owen, *Factors for One-Sided Tolerance Limits and for Variables Sampling* Plan, Sandia Corporation Monograph SCR-607, Sandia National Laboratories, Albuquerque, NM, March 1963.

7. J.C. Helton, *Quantification of Margins and Uncertainties: Conceptual and Computational Basis*, Reliability Engineering & System Safety, Volume 96, Issue 9, September 2011.

8. J.C. Helton and J.D. Johnson, *Quantification of Margins and Uncertainties: Alternative Representations of Epistemic Uncertainty*, Reliability Engineering & System Safety, Volume 96, Issue 9, September 2011.

9. B. Rutherford, V. Romero, and J. Castro, *QASPR Project FY12 QMU Milestone*, Sandia National Laboratories, Albuquerque, NM, September 2012.

DISTRIBUTION

1	MS0428	J.R. Fellerhoff	00400
1	MS0482	J.F. Nagel	00410
1	MS0829	J.M. Sjulin	00413
1	MS0482	M.L. Hoover	00420
1	MS0482	T.D. Woodall	00430
1	MS0830	C.A. Wenner	00431
1	MS0829	S.V. Crowder	00431
1	MS0830	S. Fitchett	00431
1	MS0829	L.L. Halbleib	00431
1	MS0830	J.W. Lane	00431
1	MS0829	J.T. Newcomer	00431
1	MS0829	B.M. Rutherford	00431
1	MS0829	E.V. Thomas	00431
1	MS0110	K.F. Alvin	01220
1	MS1318	T.G. Trucano	01440
1	MS0828	M. Pilch	01514
1	MS0386	T.S. Edwards	01527
1	MS0828	R.G. Hills	01544
1	MS0447	F.M. Cranfill	02111
1	MS0483	N. Brannon	02112
1	MS0483	D. Wartman	02112
1	MS0481	S.E. Klenke	02115
1	MS0457	R.A. Paulsen	02210
1	MS0878	D.S. Walsh	02735
1	MS0639	C.A. Cruz	02950
1	MS0639	K.V. Diegert	02950
1	MS0634	J.F. Lorio	02951
1	MS0634	B.J. Gomez	02957
1	MS0635	L.M. Hickman	02957
1	MS0635	G.E. Lunsford	02957
1	MS0633	N.M. Zenker	02957
1	MS0757	K.W. Mitchiner	06612
1	MS9007	R.M. Zurn	08225
1	MS9007	R.L. Bierbaum	08225
1	MS9007	A.A. Robertson	08225
1	MS9014	D.M. Skala	08238
1	MS9042	M.E. Gonzales	08250
1	MS0899	Technical Library	09536 (electronic copy)