Bayesian Modeling and Test Planning for Multi-phase Reliability Assessment

Kassandra M. Fronczyk
Lawrence Livermore National Laboratory
Livermore, CA
fronczyk1@llnl.gov

Alyson G. Wilson
North Carolina State University
Raleigh, NC
agwilso2@ncsu.edu

James F. Gilman
North Carolina State University
Raleigh, NC
jfgilman@ncsu.edu

December 15, 2017

Abstract

We use a Bayesian hierarchical model to assess the reliability of the Joint Light Tactical Vehicle (JLTV), which is a family of vehicles. The proposed model effectively combines information across three phases of testing and across common vehicle components. The analysis yields estimates of failure rates for specific failure modes and vehicles as well as an overall estimate of the failure rate for the family of vehicles. We are also able to obtain estimates of how well vehicle modifications between test phases improve failure rates. In addition to using all data to improve on current assessments of reliability and reliability growth, we illustrate how to leverage the information learned from the three phases to determine appropriate specifications for subsequent testing that will demonstrate if the reliability meets a given reliability threshold.

Keywords: Assurance Testing, Combining Information, Defense Acquisition, Reliability, Reliability Growth

1 Introduction

Reliability is a high priority in the testing of military systems. A common interest is *reliability growth*, which tracks the change, and ideally improvement, in the reliability of a system as it moves through testing phases and undergoes periodic corrective actions. Ideally, plans for future test events should be based on what is observed in previous testing.

In this paper, we look at combining information across multiple test phases with the objective of planning a future test. When combining information, there is no omnibus solution. Rather, models need to be carefully considered and evaluated to ensure that they accurately reflect the data and the underlying physical processes. Our analysis is motivated by reliability data from the Joint Light Tactical Vehicle (JLTV), which is a family of vehicles designed to replace one-third of the legacy high-mobility multipurpose wheeled vehicle (Humvee) fleet. There are four unique types of vehicles, and within types there are many possible variants. The primary mission of the family of vehicles is to provide ground mobility that is deployable worldwide and capable of operating across the range of military roles including combat, sustainment, police action, peacekeeping, and security patrol, in all weather and terrain conditions.

The JLTV has been through a series of test events as it has been developed. Engineering and Manufacturing Development (EMD) included three phases of testing, where fixes occurred only during a set Corrective Action Period (CAP) between each phase. For every vehicle, each failure encountered during testing was recorded and attributed to a specific failure mode. Since fixes were delayed to CAPs, analysis generally shows a distinct jump in the system reliability between test phases.

We propose a model that can effectively combine information across the three phases of testing and across common vehicle components. We use a Bayesian hierarchical model to assess the reliability of the family of vehicles. The model accounts for the commonality across vehicle types and allows for uncertainty quantification. Inferential objectives from the proposed model include an estimate of the mean miles between failure (MMBF) for each vehicle, failure mode, and phase of test, and an assessment of the effectiveness of the corrections completed in the CAPs. Additionally, because future tests are being planned, we propose methodology to leverage the information learned from the three phases to determine the length of testing needed to demonstrate a given reliability threshold in subsequent testing.

2 Data

Eight vehicle prototypes were used for reliability, availability, and maintainability (RAM) testing. There were four two-seat utility vehicles, two four-seat general purpose vehicles, a two-seat heavy guns carrier, and a two-seat close combat weapons carrier. The close combat weapons carrier, one general purpose, and two utility vehicles were tested in Aberdeen, Maryland, where it is much colder in the winter months but the terrain is not as harsh, and the heavy weapons carrier, one general purpose, and two utility vehicles were tested in Yuma, Arizona, where it is warmer but with more challenging terrain, to attain testing in different environmental and terrain conditions. The vehicles were driven roughly 3000 to about 5000 miles in Phase 1, between 4500 and 7500 miles in Phase 2, and between 5000 and 6700 miles in Phase 3.[1]

[1]The exact mileage has been altered to protect proprietary information.

There are many similarities between vehicle types; in fact, the most dissimilar vehicles have over 80% common parts. Consequently, some failure modes, such as brakes and radios, will be common to all eight vehicles. Other failure modes will be related, but not identical, as with hydraulics, frame, and body.

There are 1412 observed failures attributed to 26 failure modes recorded across all eight vehicles and all three test phases. In any given phase of test, every vehicle is not guaranteed to have a failure of all 26 failure modes, and each test phase does not necessarily end with a failure. Therefore, we have 624 right censored observations, accounting for miles driven without observed failures.

3 Methodology

3.1 Modeling Reliability

A standard reliability analysis employed by the Department of Defense (DoD) test community considers each test phase independently and uses the exponential distribution to model the miles between failure for the system [1]. The traditional analysis is overly simplistic, relies on correct modeling assumptions and ignores valuable information learned about the individual vehicles and their failure modes.

In this paper, we propose an alternative approach using a Bayesian hierarchical model. The methodology section is organized as follows. In section 3.1 we begin by introducing a general structure that models the relationships in the data across all test phases while accounting for CAPs and incorporates known similarities between vehicle failure modes. Next, we will consider different distributional assumptions and their implications. Using the JLTV case study we will illustrate the process of model selection using three different diagnostic checks. In section 3.2 we discuss methods of designing next stage test plans. First we outline the traditional approaches, then we walk through a Bayesian approach that allows the practitioner to incorporating the models developed from section 3.1 with the hopes of reducing testing resources required while still minimizing producer and consumer risks.

In this reliability setting we are analyzing time of failure data measured in miles for a multiple component system with the primary objective being assessment of the system failure rate and prediction of future failure counts. The most common models used in this failure rate analysis are lifetime distributions and the most common for failure count analysis is the *Poisson process* family of counting model. We will utilize both and specifically the relationships between them in planning future test.

3.1.1 Model Structure

Introducing notation, for each phase of testing we have failure mileage data for a given vehicle i and failure mode j denoted as $t_{ij0}, t_{ij1}, ..., t_{ijn_{ij}}, t_{ijc}$, where $t_{ij0} = 0$ and t_{ijc} is the censored observation at the end the testing phase. In each phase of testing, we assume the vehicle miles between failures, $y_{ijk} = t_{ijk} - t_{ijk-1}$, follow a lifetime distribution $f(\cdot)$, with an unknown set of parameters that include a rate parameter. For the first phase the rate parameter for vehicle i and failure mode j will be denoted as λ_{ij}. Then for the next stage we will use the same λ_{ij} but add a multiplicative correction parameter ρ_j^{P2} that represents the change in this rate due to the CAP. This procedure is then repeated for each subsequent stage. The correction

parameter will be discussed more in the next section but the general model is as follows,

$$\begin{aligned}
&\text{Phase 1: } y_{ijk} \mid \lambda_{ij} \sim f(\cdot), \quad i = 1, 2, ..., v \quad j = 1, 2, ..., s \quad k = 1, 2, ..., n_{ij} \\
&\text{Phase 2: } y_{ijk} \mid \lambda_{ij}\rho_j^{P2} \sim f(\cdot) \\
&\text{Phase 3: } y_{ijk} \mid \lambda_{ij}\rho_j^{P2}\rho_j^{P3} \sim f(\cdot)
\end{aligned}$$

where v represents the number of vehicles, s is the number of failure modes, and n_{ij} are the number of failures of vehicle i failure mode j.

For the model we assume a prior distribution on failure rate parameter λ_{ij}. This distribution depends on whether failure mode j is considered to be common across vehicles or related but not identical. If failure mode j is considered common across vehicles we will assume there is a single failure rate λ_j that is shared by all eight vehicles with a prior distribution that has positive support. In the JLTV study we will use a non-informative gamma distribution. For failure modes that are considered related across vehicles but not expected to be exactly the same, we place a prior distribution on the collection of λ_{ij}; in other words, we assume each of the vehicles has a distinct failure rate for failure mode j but they arise from a common distribution. This is the common Bayesian hierarchical modeling structure, where we then place hyperprior distributions on the shared distribution's parameters. In the JLTV study we use a gamma distribution with non-informative hyperpriors for all related failure modes. Many times in practice it is difficult to know ahead of time if individual components should be modeled as common or related. In the JLTV study we fit models for each component with both model structures and used diagnostic tools that are discussed in a later section to make our final decision.

3.1.2 Fix Effectiveness

After the first CAP, Test Phase 2 begins with the repaired vehicles. To capture these revisions, the PM2 reliability growth model [2] is often used. This model explicitly captures testing phases, choices about which failure modes to correct, and the potential of not completely eliminating a failure upon repair. One of the downsides of PM2 is that many parameters of potential interest, such as the Fix Effectiveness Factor (FEF), which measures how much repairs improve failure rates, are typically fixed. A common value for FEF is 0.70. We follow the premise of this type of model, but allow a more flexible and data-driven result that is less dependent on hard-coded assumptions.

One normal assumption used in reliability growth modeling is nondecreasing failure rates; that is either the fixes were effective or had no effect, but did not degrade the family of vehicles. This should generally be the case, but because we are dealing with complex systems and testing conditions are rarely completely consistent, we will sometimes see decreases in failure rates after adjustments are made. Thus we will choose a prior distribution for the correction parameter ρ_j^{P2} that has positive support. If ρ_j^{P2} is less than one, this represents an improvement in reliability. And the same thinking is applied to ρ_j^{P3} for the correction period between phases two and three. With this structure we can use the posterior distributions of the correction parameters to make inference on the effectiveness of the repairs and the join posterior distributions to evaluate the overall system reliability within each phase.

3.1.3 Distributional Assumption

The next modeling choice that must be made is selecting an appropriate probability model This includes a selecting a sampling distribution for the data and prior distributions for each parameter on which the sampling distribution depends. For lifetime data there are a number sampling distribution to choose from. Some common choices are exponential, Weibull, lognormal, gamma, inverse Gaussian, and normal failure time models. Because all of these distributions have rate parameter (sometimes called scale = 1/rate) they can be applied to the proposed modeling structure. For this paper, we will focus on the exponential and Weibull distributions.

First the exponential model,

$$f(y_{ijk}|\lambda_{ij}) = \lambda_{ij} \exp(-\lambda_{ij} y_{ijk})$$

This is by far the most common parametric distribution used in reliability modeling because of its desirable math mathematical properties and simple interpretations. One of the convenient properties, because we are assuming independence between failure modes, is the overall system failures also follow an exponential distribution with a system failure rate $\lambda_S = \sum_j \lambda_j$. A second, possibly less attractive property, is that the hazard rate is constant over miles. For the JLTV, setting this would assume that a given component is equally likely to fail in the first mile of testing as it is in the thousandth. This may be reasonable for some reliability settings, but in most setting we observe either a decreasing hazard rate where failures are likely to occur early in testing but less likely as the test goes on, or increasing hazard rate where we see the system wear down as the test progresses.

Despite its common uses the assumption of a constant hazard rate over time is rarely justifiable. It has been well documented the issues that can arise when this assumption is violated [3]. Therefore we will also consider more flexible distributions when analyzing the data. In our analysis of the JLTV data we fit the same hierarchical model structure using the Weibull as our lifetime distribution $f(\cdot)$. Now we assume that the failure mileage t_{ijk} each follow a $Weibull(\lambda, \gamma)$ distribution with scale parameter λ and shape parameter γ, with probability density function,

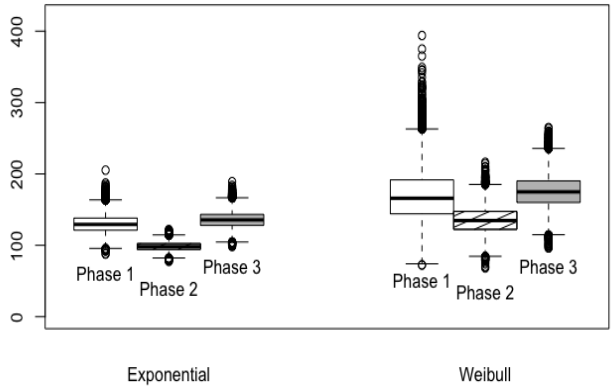
$$f(t_{ijk}|\lambda_{ij}, \gamma_i) = \lambda_{ij}\gamma_i t_{ijk}^{\gamma_i - 1} \exp(-\lambda_{ij} t_{ijk}^{\gamma_i}).$$

The exponential is a special case of the Weibull, when $\gamma = 1$. You should notice here we are now modeling the actual failure mileage t_{ijk} instead of the miles between failures y_{ijk}. This implies the "bad as old" assumption for intra-test repairs. We could have continued with the miles between failures and assumed "good as new" repairs, but "bad as old" modeling will be more convenient for developing assurance test plans in later sections. This distinction is important in the Weibull case because the failure rate is not constant in time. In practice the decision on intra-test repair assumptions should be made based on the application. In many cases repairs should be treated as somewhere between "good as new" and "bad as old". This requires more complex modeling that is beyond the scope of this paper.

It should also be noted that we are currently indexing both ρ and γ, only by j and not i, in other words, we are assuming a single shape and between phase adjustment parameter for each failure mode across vehicles. All of these parameters could also be indexed by i and modeled hierarchical or with some combination of both. This is something we will omit for this paper but hope to explore in the future.

3.1.4 JLTV Reliability

We have discussed different models structures and distributional assumptions that accompany them. Using McMC we fit each of these models to the JLTV dataset.[2] Then we use the posterior distributions of the systems parameters to estimate reliability-related quantities of interest for each of the eight vehicle types. We are able to look at a number of different statistics related to reliability at both the component and system level. In this study, one quantity of interest is expected miles to failure(EMTF). In figure 1 we have box plots and summary statistics from the posterior distributions of the expected miles to failure for one of the vehicle types for each of the three phases, first using the exponential model and second using the more general Weibull model.



Summary Statistics

Model	Phase	Mean	90% CI Low	90% CI Up
Expo	P1	129.98	110.22	152.23
Expo	P2	98.44	89.00	108.38
Expo	P3	135.89	117.78	155.45
Weibull	P1	169.96	116.95	236.16
Weibull	P2	135.24	104.89	167.48
Weibull	P3	175.51	138.87	214.23

Figure 1: Comparing EMTF for Different Models

You can see the posterior EMTF distributions from the Weibull model have more spread than the exponential. This is generally expected, given the Weibull model is estimating two model parameters to the exponentials one, but what is concerning is that the center of the distributions is also quite different between the two models. The Weibull EMTF posterior means are consistently larger than the exponentials. We saw this across all eight vehicle types. This was the first indication that the exponential model may not be a good fit to the data.

3.1.5 Model Diagnostics

As seen in the last section with the JLTV data, the decision of which model to use can have a drastic impact on reliability assessment. We will now look at a few goodness-of-fit methods that can be used to help in this decision. The first model selection question we will explore is the parametric form, exponential versus Weibull. This an example of the classic statistical modeling the trade-off between fit and interpretability.

[2]The likelihoods and sampling distributions are available in the appendix.

Because the Weibull is a more flexible model it will always, in a sense, fit the data better, but this comes at a price. The exponentials convenient form makes both computation and interpretation straightforward. When the Weibull's shape parameter is introduced this advantage is lost. Thus, when the overall fit is close to the same between the exponential and the Weibull we will default to using the exponential.

The first check we used in decide between the exponential and Weibull models is to fit the Weibull model and look at the posterior distributions of the shape parameters and determine if one is a reasonable value. Looking at the plot in figure 2 from the Weibull model fit with the JLTV data, we see that one is not a reasonable value for any of the failure modes. This is another clue that the exponential model is not be a good fit to this data.

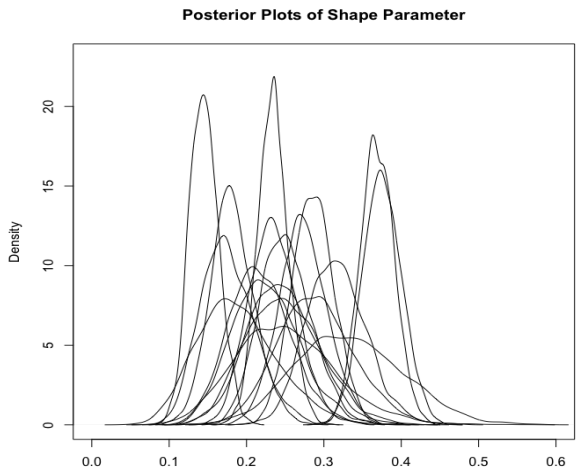


Figure 2: Posterior Plots of Weibull Shape Parameter

Now a more formal diagnostic tool, the Deviance information criterion (DIC) [4]. This is a popular method for comparing the goodness of fit of multiple models. The DIC method includes a penalty term that grows with the numbers of unknown parameters in the model. The DIC value is a unit-less measure with lower values indicating a better fit to the data. In the table below we show the DIC results for the different distribution and structure combinations, we considered with the JLTV dataset.

Goodness of Fit

Distribution	Structure	DIC
Exponential	Common Rate Model	21,264.19
	Hierarchical Rate Model	21,244.47
Weibull	Common Rate Model	20,658.25
	Hierarchical Rate Model	**17,908.06**

Table 1: JLTV DIC Results

The results in table 1 shown that the hierarchical Weibull model fits the data much better than the

exponential, but in general information criteria comparisons do not tell us if the model is truly a good representation of the underlying process. The last model diagnostics method we present is called posterior predictive checking [5]. Here we start with important features of the data that we would expect to see if the experiment was repeated, these are usually provided by practitioner with knowledge about the random process being studied. In the JLTV case one feature we were interested in was the number of miles between failure observations that were less than 140 miles. We then used the final models to simulate 5,000 new datasets for vehicles tested over the same number of miles for each of the three phases. We then plot the resulting distributions for each of the 8 vehicles for all 3 phases. In Figure 3 are two examples of the boxplots produced, one from the exponential model (left) and one from the Weibull model (right), with the dashed line showing where the value from the observed dataset fell. For this method we dont expect all of the true values to fall in the center of the distribution. Some values in the tales are to be expected in a random processes like this one, but when we see many observations in the tails, as we do with the exponential, we should consider other modeling options.

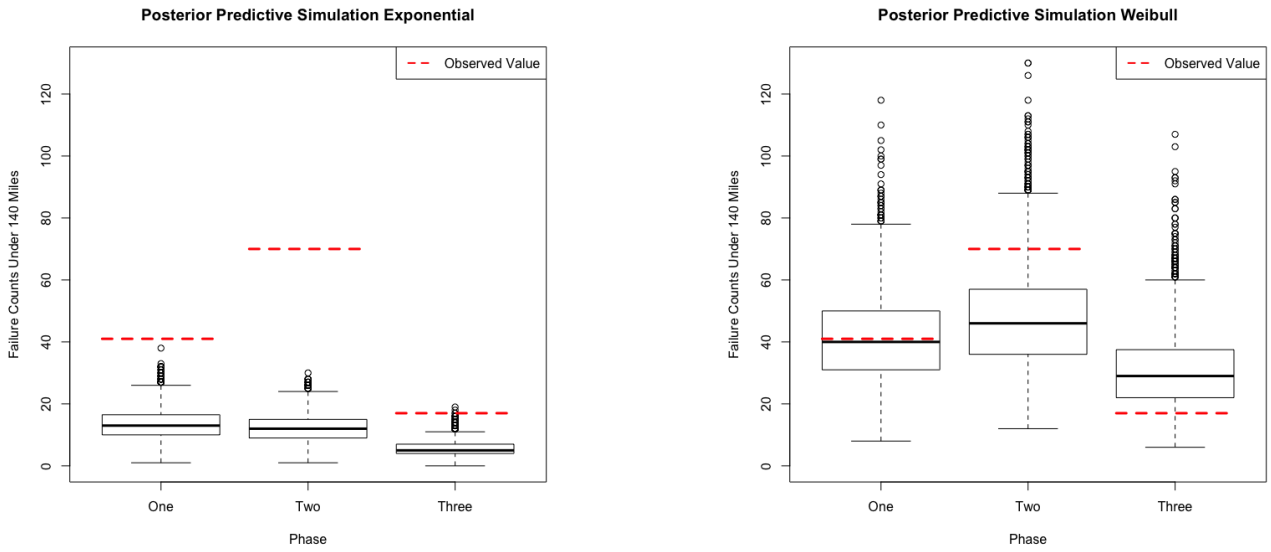


Figure 3: Posterior Predictive Plots
Exponential vs. Weibull

3.2 Assurance Test Planning

In this section, we will use the reliability models fit in section two to develop a plan for future testing. Here the objective is to demonstrate, at a desired level of confidence, that the system will meet or exceed a specified requirement. The methods to be employed will be similar to the assurance testing as discussed by Hamada et al. [6]. Bayesian *reliability assurance tests* are used to ensure that the reliability of an item meets or exceeds a specified requirement with a desired probability. Meeker and Escobar [7] established the difference between this and the more traditional *reliability demonstration test*. The difference being

that the assurance test plan incorporates supplementary data and information, where the demonstration test only uses test data.

In most test planning settings we are interested in controlling error rates while minimizing the resources required for testing. In the DoD acquisition process, the error rates are referenced in terms of risk. *Consumer risk* which considers the event of purchasing a product that does not meet reliability requirements. *Producer risk* which considers the event of a product with acceptable reliability failing a given test and not being considered for purchase. Introducing the notation for this section, in the JLTV case study, we will be determining how many miles on test T, are required and how many system failures to allow c, before the product is considered unacceptable. We will define $W(t)$ as a random variable that represents the total number of system failures after t miles. We will let π denote the quantity of interest that is related to system reliability at a given time. It is common to base both classical and Bayesian test plans on two specified levels of π: π_0, an *acceptable reliability level* (ARL), and π_1, a *rejectable reliability level* (RRL), where $\pi_1 \leq \pi_0$. Although the precise definition of ARL and RRL differ between the classical and Bayesian test criteria, we use them in an equivalent way.

3.2.1 Classical Approach

In the frequentest (classical) test planning setting the risk criteria are as follows. The *producer's risk*, defined as the probability of failing the test under the null hypothesis $\pi = \pi_0$, and the *consumer's risk*, defined as the probability of passing the test under an alternative hypothesis $\pi = \pi_1$. Suppose that we specify a maximum value, α, of the producer's risk and a maximum value, β, of the consumer's risk. For test planning, these criteria become

$$\begin{aligned} \text{Producer's Risk} &= P(\text{Test Is Failed } |\pi_0) \\ &= P(W(t) > c|\pi_0) \leq \alpha \end{aligned}$$

and

$$\begin{aligned} \text{Consumer's Risk} &= P(\text{Test Is Passed } |\pi_1) \\ &= P(W(t) \leq c|\pi_1) \leq \beta \end{aligned}$$

To choose a test plan for specified values of $(\alpha, \pi_0, \beta, \pi_1)$, we assume a distributional form that defines the relationship between the number of system failures $W(t)$ and the reliability π. Then we can simply find the combinations of these probabilities by simultaneously solving these two equations. In the case where the component failure times are assumed to be exponentially distributed with rate parameter λ_i the failure counts for each component come from a homogeneous Poisson process $w_i(t) \sim \text{Poisson}(\lambda_i t)$, and the system failure counts $W(t) \sim \text{Poisson}(\lambda_S t)$. Numerous textbooks provide additional details of this classical test planning approach.

3.2.2 Bayesian Approach

The classical methods for determining the testing procedure rely only on distributional assumptions or asymptotic results. With a Bayesian approach we incorporate the supplementary data from the previous testing phases with the hopes of minimizing the resources needed for testing.

We now consider fully Bayesian posterior risks that convey a completely different outlook from the corresponding classical risks. While the classical provide assurance that satisfactory devices will pass the test and that unsatisfactory devices will fail it, posterior risks provide precisely the assurance that practitioners often desire: if the test is passed, then the consumer desires a maximum probability β that $\pi \leq \pi_1$. On the other hand, if the test is failed, then the producer desires a maximum probability α that $\pi \geq \pi_0$. Unlike the average risks, these posterior risks are fully Bayesian in the sense that they are subjective probability statements about π.

For a test that fails, the *posterior producer's risk* is the probability that $\pi \geq \pi_0$, or $P(\pi \geq \pi_0|\text{Test Is Failed})$. Notice that this is simply the posterior probability that $\pi \geq \pi_0$ given that we have observed more than c failures. In the exponential case, if we let π be the system failure rate λ_S and π_0 be predetermined acceptable failure rate λ_0, then using Bayes' Theorem, and assuming a maximum allowable posterior producer's risk α, an expression for the posterior producer's risk for the exponential test plan (T, c) is

$$\begin{aligned}
P(\lambda_S \geq \lambda_0 \mid \text{Test Is Failed}) &= P(\lambda_S \geq \lambda_0 \mid W > c) \\
&= \int_{\lambda_0}^{\infty} p(\lambda_S \mid W > c) d\lambda_S \\
&= \int_{\lambda_0}^{\infty} \frac{f(W > c|\lambda_S)}{\int_0^{\infty} f(W > c|\lambda_S) d\lambda_S} d\lambda_S \\
&= \frac{\int_{\lambda_0}^{\infty} 1 - [\sum_{W=0}^{c} \frac{(\lambda_S T)^W exp(-\lambda_S T)}{W!}] p(\lambda_S) d\lambda_S}{\int_0^{\infty} 1 - [\sum_{W=0}^{c} \frac{(\lambda_S T)^W exp(-\lambda_S T)}{W!}] p(\lambda_S) d\lambda_S} \leq \alpha
\end{aligned}$$

For simplicity we if fix c to be zero and then perform Monte Carlo integration using N posterior draws $\lambda_S^{(j)}$

$$\begin{aligned}
P(\lambda_S \geq \lambda_0 \mid W = 0) &= \frac{\int_{\lambda_0}^{\infty} 1 - exp(-\lambda_S T) p(\lambda_S) d\lambda_S}{\int_0^{\infty} 1 - exp(-\lambda_S T) p(\lambda_S) d\lambda_S} \\
&\approx \frac{\sum_{j=1}^{N} 1 - exp(-\lambda_S^{(j)} T) I(\lambda_S^{(j)} \geq \lambda_0)}{\sum_{j=1}^{N} 1 - exp(-\lambda_S^{(j)} T)}
\end{aligned}$$

Similarly, given that the test is passed, the *posterior consumer's risk* is the probability that $\pi \leq \pi_1$, or $P(\pi \leq \pi_0|\text{Test Is Passed})$. Now we let π_1 be a predetermined rejectable failure rate λ_1 and assuming a maximum allowable posterior consumer's risk β, an expression for the posterior producer's risk for the exponential test plan (T, c) is

$$
\begin{aligned}
P(\lambda_S \leq \lambda_1 \mid \text{Test Is Passed}) &= P(\lambda_S \leq \lambda_1 \mid W \leq c) \\
&= \int_0^{\lambda_1} p(\lambda_S \mid W > c) d\lambda_S \\
&= \int_0^{\lambda_1} \frac{f(W > c|\lambda_S)}{\int_0^{\infty} f(W > c|\lambda_S) d\lambda_S} d\lambda_S \\
&= \frac{\int_0^{\lambda_1} [\sum_{W=0}^{c} \frac{(\lambda_S T)^W exp(-\lambda_S T)}{W!}] p(\lambda_S) d\lambda_S}{\int_0^{\infty} [\sum_{W=0}^{c} \frac{(\lambda_S T)^W exp(-\lambda_S T)}{W!}] p(\lambda_S) d\lambda_S} \leq \beta
\end{aligned}
$$

For simplicity we if fix c to be zero and then perform Monte Carlo integration using N posterior draws $\lambda_S^{(j)}$

$$
\begin{aligned}
P(\lambda_S \leq \lambda_1 \mid W = 0) &= \frac{\int_0^{\lambda_1} exp(-\lambda_S T) p(\lambda_S) d\lambda_S}{\int_0^{\infty} exp(-\lambda_S T) p(\lambda_S) d\lambda_S} \\
&\approx \frac{\sum_{j=1}^{N} exp(-\lambda_S^{(j)} T) I(\lambda_S^{(j)} \leq \lambda_1)}{\sum_{j=1}^{N} exp(-\lambda_S^{(j)} T)}
\end{aligned}
$$

3.2.3 Weibull Case

In the exponential case, we defined our quantity of interest that is related to system reliability at a given time π to be equal to the system failure rate λ_S. This distributional assumption gave us a convenient distributional form for our system failure counts with $W(t) \sim \text{Poisson}(\lambda_S t)$ When we assume the miles to failure for each system component follows a Weibull distribution the system failure counts no longer follow a homogeneous Poisson process. When assuming "bad as old" intra-test repairs we do find that the failure count for a given failure mode does follow a nonhomogeneous Poisson process, $w(t)_i \sim \text{Poisson}(\lambda_i t_i^{\gamma_i})$. The system failure counts then, for a given mileage t become $W(t) \sim \text{Poisson}(\sum_i (\lambda_i t_i)^{\gamma_i})$. In this case the acceptable and rejectable parameters π_0 and π_1 become quantities related to $\sum_i (\lambda_i t_i)^{\gamma_i}$. This can lead to some unpleasant integration expression, but once again we can lean on Monte Carlo integration and our evaluation for the posterior producer's risk for the Weibull test plan (T, c) becomes to

$$
\begin{aligned}
P(\sum_i \lambda_i T^{\gamma_i} \geq \pi_0 \mid W = 0) &= \frac{\int_{\pi_0}^{\infty} 1 - exp(-\sum_i \gamma_i \lambda_i T^{\gamma_i}) p(\lambda_{1,\ldots}) p(\gamma_{1,\ldots}) d\lambda_{1,\ldots} d\gamma_{1,\ldots}}{\int_0^{\infty} 1 - exp(-\sum_i \gamma_i \lambda_i T^{\gamma_i}) p(\lambda_{1,\ldots}) p(\gamma_{1,\ldots}) d\lambda_{1,\ldots} d\gamma_{1,\ldots}} \\
&\approx \frac{\sum_{j=1}^{N} 1 - exp(-\sum_i \gamma_i^{(j)} \lambda_i^{(j)} T^{\gamma_i^{(j)}}) I(\sum_i \lambda_i^{(j)} T^{\gamma_i^{(j)}} \geq \pi_0)}{\sum_{j=1}^{N} 1 - exp(-\sum_i \gamma_i^{(j)} \lambda_i^{(j)} T^{\gamma_i^{(j)}})}
\end{aligned}
$$

Where π_0 is some quantity related to the model parameters. Again we set c to be zero and used N posterior draws for both $\lambda_i^{(j)}$ and $\gamma_i^{(j)}$.

It should be noted that in this test planning process the practitioner would use the posterior distribution of the rate parameters from the final developmental testing phase $\lambda_i \rho^{P2} \rho^{P3}$. It is also common to incorporate a degradation factor to account for an expected reduction in reliability from the developmental phase to the operational phase. In the Bayesian framework, this entails adding another multiplicative parameter to the rate parameter with an appropriate prior. A 10-30 percent reduction from the developmental phase to the operational phase is common in many DoD applications.

3.2.4 JLTV Test Plan Results

The first step in applying this method to our JLTV dataset is defining π_0, π_1, α and β. That is the ARL, RRL, acceptable producer risk and acceptable consumer risk respectively. With input from subject matter experts we will define the producer's target or ARL as the expected number of failures in the first 140 miles to be less than 3, and the consumer's minimum requirements or RRL as the expected number of failures in the first 80 miles to greater than 3. We will use traditional values of 0.1 and 0.05 for α and β respectively, but in practice these values should always be set based on the true risk tolerances of the consumer and producers. This gives us the following probability statements to satisfy when designing the test:

$$\text{Consumer Risk} : P(\sum_i 80^{\gamma_i} \lambda_i \geq 3 \,|\text{Test Is Passed}) \leq 0.1$$

$$\text{Producer Risk} : P(\sum_i 140^{\gamma_i} \lambda_i \leq 3 \mid \text{Test Is Failed}) \leq 0.05$$

From here we are able to find combinations of failures allowed and test miles required that satisfy both risk requirements. Table 2 shows the resulting combinations for the JLTV data.

JLTV Test Plan

Failures Allowed	Test Miles Required
0	5,500
1	10,100
2	17,900
3	30,800
4	51,200

Table 2: JLTV Assurance Test Plan Results

These results alone do not provide much insight into how well the method is performing. A natural comparison would be with the frequentest test plan in this setting. In general we warn against this comparison because the test plan probability constraints for each method are asking different questions. The frequentest plan controls the error rates under two different assumed settings. On the other hand, the Bayesian plan puts constraints on the error when the test is passed and failed. We would argue that in most settings the Bayesian plan controls for the error the practitioner cares about. In this case there are no reasonable combinations of allowed failures and required test miles that satisfy both producer's and consumer's risk probability statements in the frequentest test plan. Both test planning approaches

rely heavily on setting reasonable values for π_0, π_1, α and β. When these are set to unrealistic values both procedures can can result in no plan that satisfies the probability statements. In this paper we have focused on test plans for single units, but th the assurance testing methodology can be extended to find test plans that allow for a changing number of testing units. In this case you would find combinations of failures allowed and test miles required and number of vehicles tested that satisfy the risk requirements. More details and examples of this can be found in the assurance testing section of Hamada et al. [6]

3.3 Discussion

The JLTV case study results provide indications that our model is successfully capturing the failure rates across failure modes, inferring logical fix effectiveness factor distributions, and providing reliability estimates that can be leveraged for future test planning. In the traditional analysis, we do not typically look at failure rates of the failure modes, nor would we be able to leverage the commonalities across vehicles. The proposed modeling framework overcomes these limitations and allows for more realistic reliability estimates for vehicles with no observed failures. Additionally, our data-driven approach for reliability growth improves on the current growth models being used in the DoD, which are based solely on fixed quantities set by management.

While our basic approach provides reasonable initial inferences, there are several extensions to incorporate. For example, reliability estimates for vehicles with no observed failures are sensible; however, we should leverage more information across vehicles. Because the correlations induced by our current model are fairly low, potential methods for inducing more correlation between the vehicles in the related failure modes are being investigated. We also need to look more closely at the intra-phase repair assumptions. Applying repairable system methodology in the modeling and reliability assessment stage could be a good approach, but this could add a great deal of complexity to the assurance planning stage.

References

[1] Director, Defense Test and Evaluation. Test and evaluation of system reliability availability maintainability - A primer; Third Edition (1982).

[2] P. M. Ellner and J. B. Hall, An approach to reliability growth planning based on failure mode discovery and correction using AMSAA projection methodology. *IEEE Proceedings of the Annual Reliability and Maintainability Symposium* (2006).

[3] National Research Council. Statistics, Testing, and Defense Acquisition: Background Papers. *National Academies Press* (1999).

[4] D. J. Spiegelhalter, N. Best, B. Carlin, and A. Van der Linde Bayesian measures of model complexity and fit. *Royal Statistical Society*, Series B, 64: 583640 (2002).

[5] Gelman, Andrew, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. (1996) *Statistica sinica* 733-760 (1996).

[6] M. S. Hamada, A. G. Wilson, C. S. Reese and H. F. Martz. Bayesian Reliability. Springer (2008).

[7] W. Q. Meeker and L. A. Escobar. Reliability: the other dimension of quailty. *Quality Technology and Quantitative Management* 1, 1-25 (2004).

4 Appendix

Exponential Model:

$$L(\lambda_1, ..., \lambda_8, \rho^{P2}, \rho^{P3}, |Y) = \prod_{i=1}^{8} \left[\prod_{j=1}^{n_j} (\lambda_i p^* e^{-\lambda_i p^* y_{ij}}) \prod_{j=1}^{R} (e^{-\lambda_i p^* y_{ij}}) \Big(\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha-1} e^{-\beta\lambda_i} \Big) \right]$$

$$p^* = \begin{cases} 1 & y_{ij} \in \text{Phase 1} \\ \rho_j^{P2} & y_{ij} \in \text{Phase 2} \\ \rho_j^{P2}\rho_j^{P3} & y_{ij} \in \text{Phase} \end{cases}$$

R is the number of right-censored observations, and n is the number of non-censored observations.

$$\pi(\alpha, \beta, \rho^{P2}, \rho^{P3}) \propto (\alpha^{a_1-1} e^{-\alpha a_2})(\beta^{b_1-1} e^{-\beta b_2})(\rho^{P2*c_1-1} e^{-\rho^{P2} c_2})(\rho^{P3*d_1-1} e^{-\rho^{P3} d_2})$$

Full conditionals for Gibbs sampling:

$$\begin{aligned} \lambda_i|\cdot &\sim \text{Gamma}(n_i + \alpha, \sum_{j=1}^{n_i} y_{ij} p^* + \beta) \\ \beta|\cdot &\sim \text{Gamma}(8\alpha + b_1, \sum \lambda_i + b_2) \\ \rho^{P2}|\cdot &\sim \text{Gamma}(N_{1,2} + c_1, \sum y_{ij}\lambda_i + c_2) \\ \rho^{P3}|\cdot &\sim \text{Gamma}(N_2 + d_1, \sum y_{ij}\lambda_i + d_2) \\ \alpha|\cdot &\sim \text{Unknown form. Used Metropolis-Hastings.} \end{aligned}$$

N_1 and N_2 are the total non-censored observations for phases two and three and phase three, respectively.

Weibull Model (Bad as Old):

$$L(\lambda_1, ..., \lambda_8, \gamma, \rho^{P2}, \rho^{P3}, |T) = \prod_{i=1}^{8} \left[\prod_{j=1}^{n_j} (\lambda_i p^* t_{ij}^{\gamma_i - 1} e^{-\lambda_i p^* t_{ij}^{\gamma_i}}) \prod_{j=1}^{R} (e^{-\lambda_i p^* t_{ij}^{\gamma_i}}) \Big(\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha-1} e^{-\beta\lambda_i} \Big) \right]$$

$$p^* = \begin{cases} 1 & t_{ij} \in \text{Phase 1} \\ \rho^{P2^\gamma} & t_{ij} \in \text{Phase 2} \\ \rho^{P2^\gamma}\rho^{P3^\gamma} & t_{ij} \in \text{Phase} \end{cases}$$

R is the number of right-censored observations, and n is the number of non-censored observations.

$$\pi(\alpha, \beta, \rho^{P2}, \rho^{P3}) \propto (\alpha^{a_1-1}e^{-\alpha a_2})(\beta^{b_1-1}e^{-\beta b_2})(\rho^{P2*c_1-1}e^{-\rho^{P2}c_2})(\rho^{P3*d_1-1}e^{-\rho^{P3}d_2})(\gamma^{g_1-1}e^{-\gamma g_2})$$

Full conditionals for Gibbs sampling:

$$\begin{aligned}
\lambda_i|\cdot &\sim \text{Gamma}(n_i + \alpha, \sum_{j=1}^{n_i} t_{ij}^{\gamma_i} p^* + \beta) \\
\beta|\cdot &\sim \text{Gamma}(8\alpha + b_1, \sum \lambda_i + b_2) \\
\rho^{P2}|\cdot &\sim \text{Gamma}(N_{1,2} + c_1, \sum t_{ij}^{\gamma_i}\lambda_i + c_2) \\
\rho^{P3}|\cdot &\sim \text{Gamma}(N_2 + d_1, \sum t_{ij}^{\gamma_i}\lambda_i + d_2) \\
\alpha|\cdot &\sim \text{Unknown form. Used Metropolis-Hastings.} \\
\gamma|\cdot &\sim \text{Unknown form. Used Metropolis-Hastings.}
\end{aligned}$$

N_1 and N_2 are the total non-censored observations for phases two and three and phase three, respectively.