INSTITUTE FOR DEFENSE ANALYSES

Bayesian Reliability: Combining Information

Alyson Wilson
Kassandra Fronczyk

July 2016

Approved for public release;
distribution is unlimited.

IDA Document NS D-8136
Log: H 16-000993

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication

One of the most powerful features of Bayesian analyses is the ability to combine multiple sources of information in a principled way to perform inference. This feature can be particularly valuable in assessing the reliability of systems where testing is limited. At their most basic, Bayesian methods for reliability develop informative prior distributions using expert judgment or similar systems. Appropriate models allow the incorporation of many other sources of information, including historical data, information from similar systems, and computer models. We introduce the Bayesian approach to reliability using several examples, and point to open problems and areas for future work.

Copyright Notice

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-8136

Bayesian Reliability: Combining Information

Alyson Wilson
Kassandra Fronczyk

Executive Summary

One of the most powerful features of Bayesian analyses is the ability to combine multiple sources of information in a principled way to perform inference. This feature can be particularly valuable in assessing the reliability of systems where testing is limited. At their most basic, Bayesian methods for reliability develop informative prior distributions using expert judgment or similar systems. Appropriate models allow the incorporation of many other sources of information, including historical data, information from similar systems, and computer models. We introduce the Bayesian approach to reliability using several examples, and point to open problems and areas for future work, including:

- Reliability for various types of systems: on-demand with pass-fail testing (notional SDB-II data) and continuous lifetime data (viscosity breakdown times). These examples include definitions and illustrations of prior distributions, likelihood and sampling distributions, posterior distributions, and predictive distributions.
- Additional discussion of how to specify prior distributions is provided, along with brief descriptions of methods and possible resources for more complex analyses like hierarchical modeling, system reliability with subsystem or component level testing, and implementation using Markov chain Monte Carlo techniques.
- Finally, some open research areas are discussed regarding combining information across multiple tests for assessment purposes and to plan an appropriately sized follow-on test.

Bayesian Reliability: Combining Information

Alyson G. Wilson[1] and Kassandra M. Fronczyk[2]

[1]North Carolina State University
[2]Institute for Defense Analyses

August 18, 2016

Abstract

One of the most powerful features of Bayesian analyses is the ability to combine multiple sources of information in a principled way to perform inference. This feature can be particularly valuable in assessing the reliability of systems where testing is limited. At their most basic, Bayesian methods for reliability develop informative prior distributions using expert judgment or similar systems. Appropriate models allow the incorporation of many other sources of information, including historical data, information from similar systems, and computer models. We introduce the Bayesian approach to reliability using several examples and point to open problems and areas for future work.

1 Background

This is an interesting time for statistical reliability. One one hand, shrinking budgets in areas like defense acquisition lead for calls to “do more with less” and “use all available data” (NRC, 1998; NCR, 2004; NRC, 2006; NRC, 2015). On the other hand, we are also in the era of “big data,” where information from sensors, warranty claims, and field data can be used to supplement traditional reliability testing (Meeker and Hong, 2014). What these challenges have in common are the need to combine multiple sources of information from different sources, (e.g., life tests, physics-based knowledge, expert opinion, computer experiments) using models that acknowledge the differences in the variation and uncertainty among the sources (Anderson-Cook, 2009; Reese et al., 2004). Bayesian statistical approaches can provide a natural and principled way to combine the information.

At their core, Bayesian methods start with Bayes’ Theorem,

$$\pi(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta)\pi(\theta)}{f(\mathbf{y})} \, . \tag{1}$$

The left-hand side of the equation is the *posterior distribution*, which summarizes the current state of knowledge about the parameters in a statistical model, given the observed data. The first term on the right-hand side of the equation is $f(\mathbf{y} \mid \theta)$, which is the *likelihood* (the distribution for the data thought of as a function of θ). The second term, $\pi(\theta)$ is the *prior distribution* for θ, which captures our state of knowledge about the parameters before observing the current data. The denominator, $f(\mathbf{y}) = \int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta$, is the marginal distribution for the data. We frequently do not compute $f(\mathbf{y})$ explicitly, since we know the posterior distribution is a probability density that integrates to 1. A good way to remember Bayes' Theorem: the posterior is proportional to the likelihood times the prior.

Bayesian methods for reliability start from Eq. 1. When we refer to a Bayesian *model*, we mean the specification of both the likelihood and the prior distribution. As with non-Bayesian approaches, much attention is paid to specifying the likelihood. While there is considerable overlap in the likelihoods considered in Bayesian and non-Bayesian reliability methods, hierarchical models (Section 3.2) and models for multi-level system reliability (Section 3.3) are more commonly discussed in a Bayesian context and are described here in some detail.

The prior distribution (Section 3.1) is a key component for Bayesian methods. There are two necessary features when using a prior distribution: (1) there is previous information relevant to the analysis, (2) this information can be summarized as a probability distribution on parameters that are useful in the current analysis. However, the situation where the analyst wants to summarize "no prior knowledge" can also be captured. Establishing a prior distribution clearly requires careful thought and modeling, but has the opportunity to supplement the data in the current experiment and the potential to provide improvements in the precision of estimates. As with all statistical modeling, the final results of an analysis using a prior distribution must be carefully examined to determine the sensitivity and impact of assumptions and modeling choices (Gelman et al., 2013; Reese et al., 2001).

Bayesian methods can also provide computational simplifications when fitting complex models. Specifically, in reliability problems, censored data can be incorporated in a very straightforward way (Section 2.2). In addition, when framed as a Bayesian problem, complex models can often be relatively easily fit using Markov chain Monte Carlo (Section 3.4). In addition, Bayesian methods easily allow the computation of distributions (to include point and interval estimates) for complicated functions of model parameters (e.g., predictions, probability of failure, quantiles of lifetime distribution), which can support additional modeling to combine information.

2 Basics

2.1 Binomial Example

Systems developed and deployed by the Department of Defense (DoD) undergo a variety of test events that help understand reliability (NRC, 1998). The company building the system should use "design for reliablity" practices (Rhoads, 2011) and contractor testing to make an initial assessment of reliability. The government performs developmental testing, which focuses on requirements checking, and operational testing, which considers the system in realistic settings and environments (Dickinson et al., 2015). During a system's lifecycle, there may be several variants that result from repairs, upgrades, or life extension programs. Ideally, we would like to design a full suite of tests for each variant of the system under all operational conditions. In practice, this is seldom possible, due to a variety of constraints (e.g., cost, time, treaty restrictions). Consequently, the problem of interest is how we use all of the information we have collected to understand the current reliability of the stockpile of systems.

As an example, consider the Small Diameter Bomb-II (SDB-II), which is a multipurpose bomb that consists of seven subsystems with multiple components that are tested with 14 end-to-end tests[1]. The response of interest is treated as pass/fail, successful detonation or not. Suppose that of $n = 14$ tests, SDB-II failed to detonate twice. The test data are modeled with the likelihood function, $f(\boldsymbol{y} \mid R)$. This likelihood function is the same starting point that would be used for a non-Bayesian reliability analysis. The binary test data of bomb detonations follow a binomial distribution with probability of a pass of R. That is,

$$f(\mathbf{y} \mid R) \propto R^{s}(1-R)^{n-s} \ ,$$

where $\mathbf{y}$ is the number of successful tests, s, and the number of failed tests, $n - s$.

The prior distribution of SDB-II reliability, $\pi(R)$, is constructed from previous data or expert knowledge. The prior reliabilities are captured in

[1] Data are notional.

the form of a distribution that is determined before the data are obtained. Suppose that SDB-II was previously tested and failed 3 out of 17 tests. Depending on how operationally realistic the previous testing was, we may choose to include none or all of the prior information into our prior assessment of reliability, $\pi(R)$. One approach to including this information is through a beta distribution

$$\pi(R) \propto R^{n_p p}(1-R)^{n_p(1-p)} ,$$

with p as the prior reliability estimate and $n_p \geq 0$ as the weighting factor of that prior estimate (Johnson et al., 2003). When n_p is set to 0, we do not believe that the prior data are relevant to the current test data and the prior distribution gives equal probability to all values between 0 and 1 (see the middle panel of Figure 1). As n_p increases, our confidence in the prior reliability estimate increases, and the distribution peaks around this estimate (see the left panel of Figure 1).

The posterior distribution is proportional to the product of the likelihood function and the prior distribution. The choice of the beta distribution as a prior is useful for several reasons: it is flexible enough to describe a variety of prior beliefs, it ensures that R is between $(0, 1)$, and it is the *conjugate* prior for the binomial distribution. Conjugate priors have the property that the form of the prior distribution, when combined with the likelihood, is the same as the posterior distribution.[2] Multiplying the likelihood and prior for

[2] Conjugate priors exist for many distributions. See Hamada et al. (2008) or Gelman et al. (2013) for more information.

SDB-II and rearranging, we have

$$\begin{aligned}\pi(R \mid \boldsymbol{y}) &\propto R^{s}(1-R)^{n-s}R^{n_p p}(1-R)^{n_p(1-p)} \\ &\propto R^{s+n_p p}(1-R)^{n-s+n_p(1-p)} \\ &\sim \text{Beta}(s+n_p p+1, n-s+n_p(1-p)+1).\end{aligned}$$

The choice of n_p and p will impact posterior inference for SDB-II reliability, as well as any functions thereof. If we use a non-informative or diffuse prior (here, $n_p = 0$; see middle panel of Figure 1), the analysis gives a mean of 0.81 and 95% credible interval of $(0.60, 0.96)$. Contrast this with the non-Bayesian maximum likelihood estimate of 0.86 and 95% confidence interval of $(0.57, 0.98)$, with slightly wider intervals and a higher point estimate. Uncertainty is decreased when our prior assessment matches what the data say (left panel of Figure 1). The prior data were collected in a semi-operationally realistic manner, and therefore we set $n_p = 7$, resulting in a reliability estimate of 0.85 and 95% interval of $(0.67, 0.96)$. The right panel of Figure 1 shows a prior assessment of reliability that is rather poor (failing 9 out of 17 tests) but the testing was only partially relevant to the current test and is downweighted ($n_p = 3$). These settings result in a SDB-II reliability estimate of 0.78 and a 95% credible interval of $(0.57, 0.94)$.

In this case, we have chosen a specific n_p; however, using more complex models, the weight can be chosen based on the observed data (Reese et al., 2004; Ibrahim and Chen, 2000; Anderson-Cook et al., 2007).

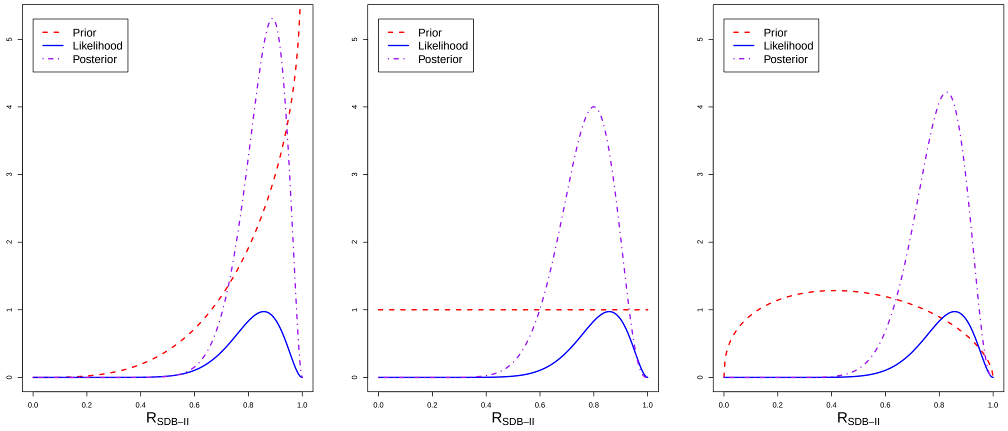

Figure 1: Prior (dashed red lines), likelihood (solid blue line), and posterior (dot-dash purple lines) distributions for the SDB-II reliability analysis with different prior settings in each panel.

A good statistical analysis should include some check of the adequacy of the model fit to the data. Sensitivity analysis investigates how much inference changes when other reasonable priors or models are assumed instead of the one in use. In the case of SDB-II, the resulting posteriors do not change drastically under each prior assessment. Depending on the purpose of the analysis, posterior predictive checking (see Section 2.2) can help determine the adequacy of model fit and how the model is impacted by changing the prior.

2.2 Lifetime Example

Now suppose that we consider the data in Table 1. These data are the viscosity breakdown times (in 1000s of hours) for 50 samples of a lubricating

5.45	16.46	15.70	10.39	6.71	3.77	7.42	6.89	9.45	5.89
7.39	5.61	16.55	12.63	8.18	10.44	6.03	13.96	5.19	10.96
14.73	6.21	5.69	8.18	4.49	3.71	5.84	10.97	6.81	10.16
4.34	9.81	4.30	8.91	10.07	5.85	4.95	7.30	4.81	8.44
6.56	9.40	11.29	12.04	1.24	3.45	11.28	6.64	5.74	6.79

Table 1: Viscosity breakdown times (in 1000s of hours) for 50 samples of a lubricating fluid (from Hamada et al. (2008)).

fluid. Unlike the data in Section 2.1, these data are continuous and an example of *lifetime* data.

There are a variety of distributions that are commonly used when analyzing lifetime data, which include the exponential, gamma, Weibull, and lognormal. These distributions capture a variety of different features, including different hazard functions. We can think of the hazard function as an items propensity to fail in the next short interval of time, given that the item has survived to time t, and we define it as

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{R(t)}$$

An exponential distribution has constant hazard, while a Weibull distribution can have constant, increasing, or decreasing hazard, depending on the choice of parameters.

One way to choose the appropriate sampling distribution for our data is to consider a sequence of probability plots for each different distribution; this suggests the lognormal as appropriate for our data. The lognormal has probability density function

$$f(x \mid \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\log(x) - \mu)^2\right], \ x > 0, \ -\infty < \mu < \infty, \ \sigma > 0$$

with

$$\begin{aligned}
\mathrm{E}(X) &= \exp(\mu + \frac{\sigma^2}{2}) \\
\mathrm{Var}(X) &= \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2) \\
h(t \mid \mu, \sigma) &= \frac{\phi\left(\frac{\log(t)-\mu}{\sigma}\right)}{\sigma t - \sigma t \varPhi\left(\frac{\log(t)-\mu}{\sigma}\right)}
\end{aligned}$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution and $\varPhi(\cdot)$ is the cumulative distribution function of the standard normal distribution. If a random variable X is log-normally distributed, then $Y = \log(X)$ has a normal distribution. The lognormal distribution has two parameters, μ and σ^2, which correspond to the mean and variance of the distribution of $\log(X)$. If σ^2 is known, then the normal distribution is the conjugate prior for μ; if μ is known, then the inverse gamma distribution is the conjugate prior for σ^2.

Suppose that we specify that $\mu \sim$ Normal$(2, 1)$ and independently $\sigma^2 \sim$ InverseGamma$(6, 5)$. A useful tool to assess the choice of prior distribution for the parameters is the prior predictive distribution

$$p(y) = \int f(y \mid \mu, \sigma^2)\pi(\mu, \sigma^2) d\mu d\sigma^2$$

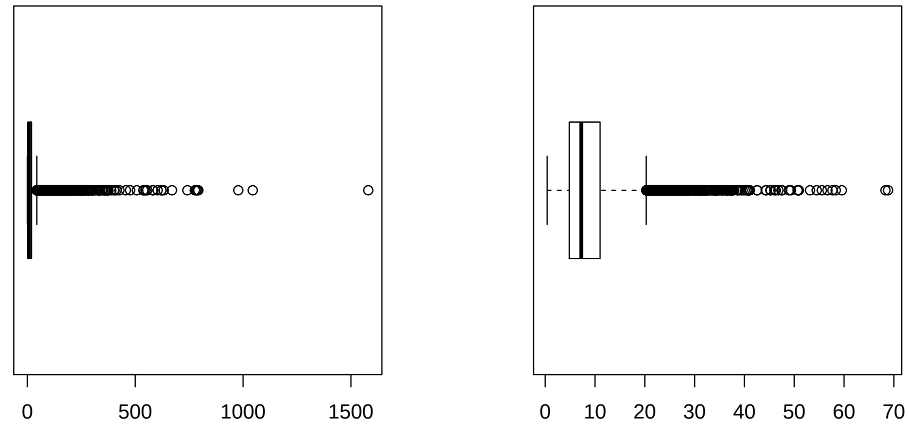

Figure 2: Prior (left) and posterior (right) predictive distributions for viscosity data with lognormal likelihood and prior distributions of $\mu \sim \text{Normal}(2, 1)$ and $\sigma^2 \sim \text{InverseGamma}(6, 5)$.

This distribution, shown in the left panel of Figure 2, reflects what we would expect for a randomly selected fluid breakdown time in the presence of all *a priori* uncertainty. Instead of performing the integration, we draw 10,000 observations from the prior, used each pair to draw an observation from a lognormal distribution, and draw a boxplot of the results.

The prior distribution for this problem is not conjugate. However, the posterior distribution is still determined using Bayes' Theorem (Eq. 1). For this problem,

$$\begin{aligned}
\pi(\mu, \sigma^2 \mid \mathbf{y}) &\propto \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\mu - 2)^2) \frac{5^6}{\Gamma(6)} (\sigma^2)^{-7} \exp(-\frac{5}{\sigma^2}) \\
&\quad \prod \frac{1}{y_i \sigma \sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(\log(y_i) - \mu)^2) \\
&\propto \frac{1}{\sigma^{64}} \exp(-\frac{1}{2}(\mu - 2)^2 - \frac{5}{\sigma^2}) \exp(-\sum \frac{(\log(y_i) - \mu)^2}{2\sigma^2}) \,.
\end{aligned}$$

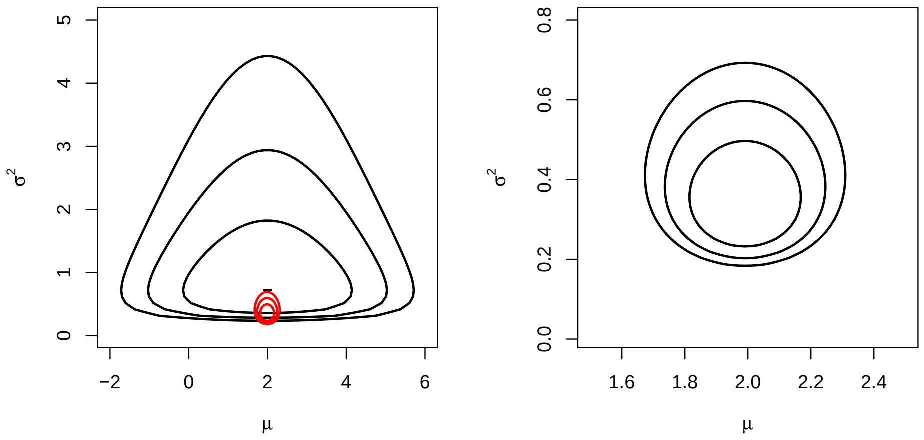

Figure 3: (Left) Contour plot of prior distribution with posterior distribution overlaid; (Right) Zoomed in on the contour plot of the posterior distribution

Figure 3 shows a contour plot of the prior distribution with the posterior overlaid in red (left) and a zoomed in contour plot of the posterior (right).

In much the same way as we calculated the prior predictive distribution, we can compute a posterior predictive distribution

$$p(x) = \int f(x \mid \mu, \sigma^2)\pi(\mu, \sigma^2 \mid \mathbf{y})d\mu d\sigma^2$$

To avoid computing the integral, we use 10,000 samples from the posterior distribution, use each sampled pair to draw an observation from a LogNormal($\mu^{(i)}$,$(\sigma^2)^{(i)}$) distribution, and draw a boxplot of the results (right panel, Figure 2). The general technique to draw samples from the posterior distribution is Markov chain Monte Carlo, which is briefly described in Section 3.4.

The posterior predictive distribution shows what we expect to see if we draw another observation. It integrates over our current *a posteriori* uncertainty about the model parameters. We can extend this idea to do model checking and see if our model is consistent with the data. The idea is that if our model fits, then replicated data generated under the model should look similar to observed data (Gelman et al., 2013). More specifically, the observed data should look plausible under the posterior predictive distribution.

The basic technique is to draw simulated values of replicated data from the posterior predictive distribution and compare some summary of these samples to the same summary of the observed data. Any systematic differences between the simulations and the data indicate potential failings of the model. In our example, we use our posterior draws, draw a replicate data set of size 50 from a LogNormal($\mu^{(i)}$,$(\sigma^2)^{(i)}$) distribution, compute a summary statistic, draw a histogram, and compare to the observed data. In Figure 4 we show two summary statistics: the deviance (-2 * log(likelihood)), which is a general measure of goodness of fit, and the 75th percentile of the data. There is no evidence of a discrepancy between the model and the observed data for these two features.

Censored data could be easily incorporated into the computation of the posterior distribution using the expressions in Table 2 in the likelihood. In Eq. 2, all of the data was uncensored, so that each observation made a contribution of $f(t)$ to the likelihood.

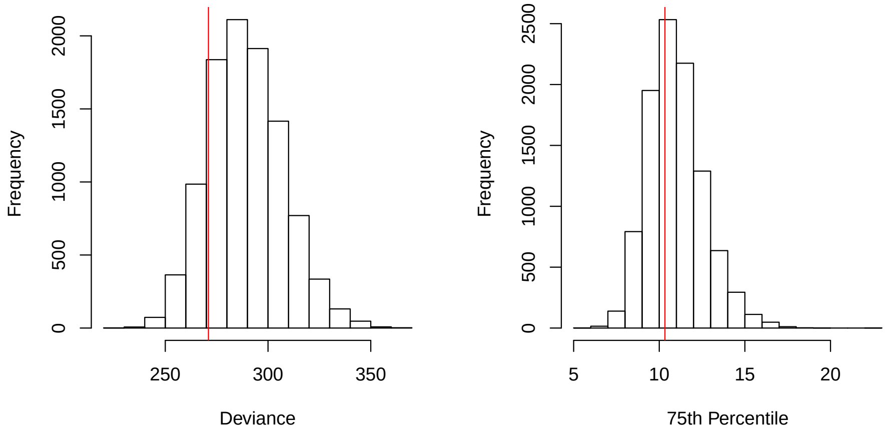

Figure 4: Histograms created from replicate data sets of size 50 drawn using posterior samples of (μ, σ^2). (Left) Computed deviance from replicate data sets, with observed data deviance in red. (Right) Computer 75th percentile from replicate data sets, with observed data 75th percentile in red.

Type of Observation	Failure Time	Contribution
Uncensored	$T = t$	$f(t)$
Left censored	$T \leq t_L$	$F(t_L)$
Interval censored	$t_L < T \leq t_R$	$F(t_R) - F(t_L)$
Right censored	$T > t_R$	$1 - F(t_R)$

Table 2: Likelihood contributions for censored data.

3 More Details

3.1 Priors Distributions

Many people are uncomfortable with the Bayesian approach, often because they view the selection of a prior as being arbitrary and subjective. The prior distribution should capture the information known about the component or system of interest and be defensible. Careful thought should always be put into the prior distribution, as naively specified priors can lead to misleading results. Building a prior begins with the properties of the parameter of interest: if a parameter needs to be positive, choose a distribution that is also positive. From there, prior construction can be broadly grouped into the specification of *informative* and *non-informative* distributions.

The non-informative prior is also commonly referred to as a flat, diffuse, vague, or objective prior. In general, a non-informative prior tries to capture the idea of minimal knowledge about the parameter. These priors include only basic information about parameters, like the reliability has a uniform chance of being any value between 0 and 1. See Berger (2006) and Ghosh (2011) for discussion and examples. Note that Jeffrey's priors are priors that are invariant under reparametrization of the parameters. While they are considered objective, these priors are not always proper (i.e., they do not integrate to 1) and may not perform satisfactorily in some cases (Box and Tiao, 1973; Datta and Ghosh, 1996).

Informative priors can be based on subject matter expert and subjective assessments (see Von Winterfeldt and Edwards (1986); Morgan and Henrion (1991); U. S. Nuclear Regulatory Commission (1994); Meyer and Booker (2001); Garthwaite et al. (2005); Bedford et al. (2006); Goldstein (2006); O'Hagan et al. (2006)), or previous test data (e.g., Johnson et al. (2005); Dickinson et al. (2015)). Some general notes on developing priors: ensure that the prior information is relevant to the current reliability evaluation. Allow for the analysis to change freely based on the data observed. Be mindful that any value of reliability with zero probability in the prior has zero probability in the posterior, regardless of the amount of data observed. It is always prudent to check impact of the prior assumptions: explore the prior predictive distributions and re-check the analysis with a sensitivity study. A good model should be fairly robust to prior specifications.

3.2 Hierarchical Models

Situations arise in reliability assessments where multiple parameters are thought to be similar but not identical. Consider the failure rate for a family of vehicles. Here, knowing that the vehicles are built on the same chassis or have common parts means data about failure rates from one vehicle variant also provides information about the failure rate of the other variants. In Dickinson et al. (2015), the authors use a hierarchical model for the Stryker family of vehicles and leverage information across vehicle variant and test phase. Hierarchical models are widely applicable and can provide insight into complex applications.

Suppose that a new torpedo is fit with wings and can be dropped from either a helicopter or a low flying airplane. We are interested in the miss distance of the torpedo (i.e., the distance from the aim point to the actual splash point in the water). Testing occurred on two helicopter variants and three types of airplanes. Due to the placement of the launchers on the various aircraft, the accuracy of the torpedo is expected to be similar but not identical depending on which launcher is used. The likelihood function for the resulting data, $\mathbf{y}$, is Normal, with mean μ and variance σ^2. The variability is determined to be constant across variant, but each launcher will have a distinct mean. The prior for the means will also be Normal with mean θ and variance τ^2 (see the model specification in 2).

$$\begin{aligned} y_{ij} \mid \mu_i, \sigma^2 &\sim \mathrm{N}(\mu_i, \sigma^2) \\ \mu_i \mid \theta, \tau^2 &\sim \mathrm{N}(\theta, \tau^2) \end{aligned} \tag{2}$$

Here, i indicates the launcher types ($i = 1, \ldots, 5$) and j denotes the tests for a given launcher type ($j = 1, \ldots, n_i$). The hierarchical model leverages information from all launcher types while still allowing for distinct mean values.

3.3 System Reliability

In the discussion so far, we have modeled systems without considering their constituent components. However, Bayesian methods are readily applicable

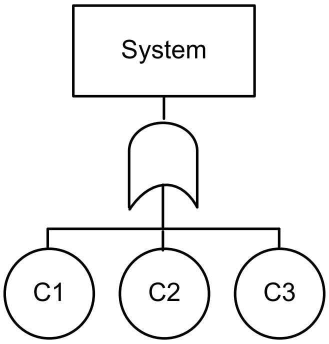

Figure 5: Three-component series system.

Table 3: Data for three-component series system with system data

	Successes	Failures	Units Tested
Component 1	8	2	10
Component 2	7	2	9
Component 3	3	1	4
System	10	2	12

to assessing the reliability of systems. Consider the fault tree in Figure 5, which is a series system where the system fails if any component fails. Suppose we have the data given in Table 3, which shows independent pass/fail data for each component and for the system as a whole.

Let R_i be the reliability for component i, $i = 1, 2, 3$. As in the example from Section 2.1, the likelihood for each component can be written as

$$L(\mathbf{y}_i \mid R_i) \propto R_i^s (1 - R_i)^{n_i - s} \ ,$$

or more concretely for this problem, we can write the likelihood for the first three rows of component data as

$$f(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3 \mid R_1, R_2, R_3) \propto R_1^8(1-R_1)^2 R_2^7(1-R_2)^2 R_3^3(1-R_3)\ .$$

To complete the specification of the likelihood, we also need to include the system data. For a series system, we know that the system reliability is equal to the product of the component reliabilities: $R_S = R_1R_2R_3$. To include the system data, we have

$$\begin{aligned} &f(\mathbf{y} \mid R_1, R_2, R_3) \\ &\quad \propto R_1^8(1-R_1)^2 R_2^7(1-R_2)^2 R_3^3(1-R_3) R_S^{10}(1-R_S)^2 \\ &\quad \propto R_1^8(1-R_1)^2 R_2^7(1-R_2)^2 R_3^3(1-R_3)(R_1R_2R_3)^{10}(1-R_1R_2R_3)^2 \end{aligned}$$

To complete the Bayesian analysis, we now specify a prior distribution on our three unknown component reliabilities. This specification must be done with considerable care. For example, Figure 6 shows the induced prior distribution when a uniform distribution is assumed for the three component reliabilities. Note that the prior is somewhat pessimistic about the prior distribution of system reliability, and this pessimism is only compounded as the number of components increases (Parker, 1972). As multi-level models with data for systems and components become more complicated, the specification of prior distributions also become increasingly difficult (see, for example, Allella

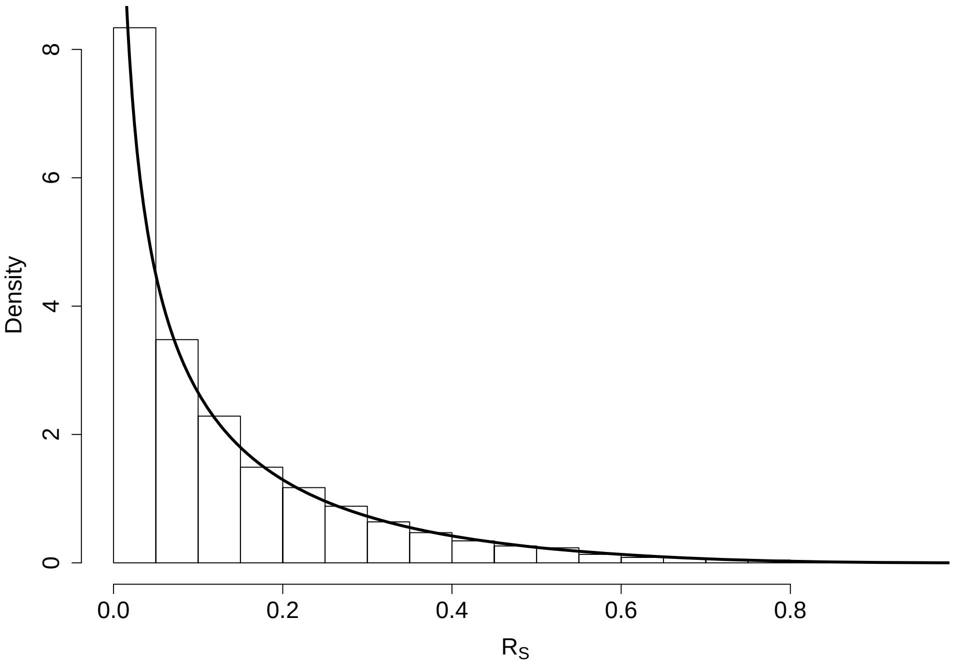

Figure 6: Induced prior distribution on three-component series system reliability with uniform prior distributions on each component reliability. The histogram comes from simulation and the solid line is actual prior density function.

et al. (2005); Zoh (2012); Guo and Wilson (2013)). Developing robust prior distributions for systems is an ongoing area of research.

Estimating reliability when no failures have been observed does not create complications for the Bayesian approach: one simply specifies a prior distribution and uses the likelihood from the observed data to get to a posterior distribution. Even for a case where there is no observed data, the Bayesian approach has a reasonable solution. For a single component with no data, the posterior is the same as the prior.

The methods that we describe here have been extended in a variety of ways. For example, Anderson-Cook et al. (2008) considers how to incorporate multiple diagnostics measured at the components; Guo and Wilson (2013) describes models for binary, lifetime, degradation, and expert opinion at the component and system level; Wilson et al. (2007) describes the development of complex system representations; Wilson and Huzurbazar (2007) generalize the system structures to Bayesian networks; Anderson-Cook (2008) and Zhang and Wilson (2016) describe model checking for incorrect system structure or dependent data.

3.4 Implementation

Many proposed Bayesian models are analytically intractable, i.e., conjugate prior distributions do not exist or do not fit physical or theoretical constraints. Unless you are working with only a few parameters, the posterior distribution is obtained by way of Markov chain Monte Carlo (MCMC) methods (Gamerman and Lopes, 2006; Albert, 2009). MCMC algorithms can be thought of as general-purpose methods to obtain samples from an arbitrary distribution – in the case of a Bayesian model, the posterior distribution. The posterior samples can be used to provide posterior estimates of the parameters of interest, as well as posterior credible intervals. Posterior samples can also be transformed to give point and interval estimates of any function of the parameters, like the hazard or survival function from a given reliability model or a combination of component reliabilities to estimate the full system reliability. For further details, see Hamada et al. (2008), Robert and Casella (2010) and Gelman et al. (2013).

Note there are many software packages that implement Bayesian models, including OpenBUGS, JAGS, SAS/PROC MCMC, and R packages (mcmc, arm, bayesSurv, rstan), and many more.[3]

4 Areas for Further Research

4.1 Combining Data Across Tests

One of the ongoing challenges for combining information in reliability is how to use information from multiple tests. Over time, the system changes (e.g., through repair or redesign) and the test environments change (e.g., developmental to operational testing). There are typically not enough resources to fully test each variant of the system during each test event, so the challenge that arises is how to combine information from the test events to characterize the system and its reliability in multiple environments. Anderson-Cook (2009, p. 241) highlights the potential advantage of solving this problem, “If we have multiple small data sets that are each individually insufficient to answer the question of interest, then combining them and incorporating engineering or scientific understanding of the process should allow us to extract more from that collection of data compared to just looking at the pieces

[3] For more tools and resources, see `https://cran.r-project.org/web/views/Bayesian.html`.

alone."

When combining information, there is no omnibus solution. At its simplest, the problem of combining information across tests involves identifying parameters (or functions of parameters) that appear in models for multiple tests. This implies that data from multiple tests provides information to estimate the parameters. However, the models need to be carefully considered and evaluated to ensure that they accurately reflect the data and the underlying physical processes. The models have to be simple enough that they can be distinguished by the data, but at the same time complex enough to capture the physical processes. One potentially promising approach is to consider a hybrid of reliability growth models (National Research Council Panel on Reliability Growth Methods for Defense Systems, 2015), that capture the arc of the test process, with models that capture, either empirically or physically, details of the individual systems.

4.2 Assurance Testing

In an era of high reliability requirements and limited resources, leveraging previous test data to plan the next test is essential. Here the objective is to demonstrate that at a desired level of confidence, the system will meet or exceed a specified requirement. Bayesian assurance tests are used to insure that the reliability of an item meets or exceeds a specified requirement with a desired probability. Although practitioners often use "assure" and "demonstrate" synonymously, Meeker and Escobar (2004) distinguish between reliability demonstration and reliability assurance testing. A *reliability demonstration test* is essentially a classical hypothesis test, which uses only the data from the current test to assess whether the reliability-related quantity of interest meets or exceeds the requirement. A *reliability assurance test*, however, uses supplementary data and information to reduce the required amount of testing.

Consider SDB-II as an example. Given previous test data on each subsystem (if available) and the 14 end-to-end tests, assurance testing ideas can be used to plan the next test phase. We want to determine (n, c) where n is the test sample size and c is the number of systems allowed to fail before the "test is failed." There are two errors we could make, either we decide the "test is failed" when SDB-II reliability R is higher than a specified π_P or decide the "test is passed" when SDB-II reliability is lower than a specified π_C. These errors are the *posterior producer's risk* (choose a test plan so that if the test is failed, there is a small probability that the reliability at t_I (the time of interest) is high) and the *posterior consumer's risk* (choose a test plan so that if the test is passed, there is a small probability that the reliability at t_I is low).

The posterior producer's risk is shown mathematically below. Looking at line (3), this is the probability that $R \geq \pi_P$ (the integrand) given everything known about R (i.e., $p(R \mid \mathbf{x})$ from Section 2.1) and that we observe more than c failures (in brackets).

$$\begin{aligned}
Posterior\ Producer's\ Risk &= \boldsymbol{P}(R \geq \pi_P \mid Test\ Is\ Failed, \mathbf{x}) \\
&= \int_{\pi_P}^{1} p(R \mid y > c, \mathbf{x}) dR \\
&= \int_{\pi_P}^{1} \frac{f(y > c \mid R) p(R \mid \mathbf{x})}{\int_0^1 f(y > c \mid R) p(R \mid \mathbf{x}) dR}\, dR \qquad (3) \\
&= \frac{\int_{\pi_P}^{1} \left[\sum_{y=c+1}^{n} \binom{n}{y} (1-R)^y R^{n-y}\right] p(R \mid \mathbf{x}) dR}{\int_0^1 \left[\sum_{y=c+1}^{n} \binom{n}{y} (1-R)^y R^{n-y}\right] p(R \mid \mathbf{x}) dR} \\
&= \frac{\int_{\pi_P}^{1} \left[1 - \sum_{y=0}^{c} \binom{n}{y} (1-R)^y R^{n-y}\right] p(R \mid \mathbf{x}) dR}{1 - \int_0^1 \left[\sum_{y=0}^{c} \binom{n}{y} (1-R)^y R^{n-y}\right] p(R \mid \mathbf{x}) dR}
\end{aligned}$$

The posterior consumer's risk is shown mathematically below. Looking at line (4), this is the probability that $R \leq \pi_C$ (the integrand) given everything known about R (i.e., $p(R \mid \mathbf{x})$ from Section 2.1) and that we observe no more than c failures (in brackets).

$$\begin{aligned}
Posterior\ Consumer's\ Risk &= \boldsymbol{P}(R \leq \pi_C \mid Test\ Is\ Passed, \mathbf{x}) \\
&= \int_{0}^{\pi_C} p(R \mid y \leq c, \mathbf{x}) dR \\
&= \int_{0}^{\pi_C} \frac{f(y \leq c \mid R) p(R \mid \mathbf{x})}{\int_0^1 f(y \leq c \mid R) p(R \mid \mathbf{x}) dR}\, dR \qquad (4) \\
&= \frac{\int_{0}^{\pi_C} \left[\sum_{y=0}^{c} \binom{n}{y} (1-R)^y R^{n-y}\right] p(R \mid \mathbf{x}) dR}{\int_0^1 \left[\sum_{y=0}^{c} \binom{n}{y} (1-R)^y R^{n-y}\right] p(R \mid \mathbf{x}) dR}.
\end{aligned}$$

With the posterior producer and consumer risks defined, the number of

SDB-II tests and allowable failures are chosen such that both risks are below a threshold. For more details and examples, see Hamada et al. (2008) or Hamada et al. (2014).

Frequently, test planning for a group of related systems requires assurance testing ideas. For instance, a family of vehicles may go through multiple phases of test but the next test will only have three of five variants available. To obtain a reliability assessment of the family, information must be leveraged across both test phase and variants. There may also be other covariates, such as test site or two-seat and four-seat configurations. These extensions to the assurance testing methodology are areas of future research.

5 Conclusion

Bayesian methods provide a principled way to combine information for reliability. They allow inferences and uncertainty quantification for complex models and are relatively easy to implement with the ever-increasing choices for software. While we have illustrated these methods using DoD systems, Bayesian approaches are applicable to a wide variety of problems.

Methods for combining data require detailed thought and analysis at every step of the process. Care must be taken to identify relevant information for prior development, specify the likelihood, understand the model structure (e.g., for a system or a hierarchical specification), check the sensitivity to assumptions, and examine model fit. This is good statistical practice for any analysis involving complex models.

References

J. Albert. *Bayesian Computation with R.* Springer, New York, NY, 2nd edition, 2009.

F. Allella, E. Chiodo, D. Lauria, , and M. Pagano. Optimal reliability allocation under uncertain conditions, with application to hybrid electric vehicle design. *International Journal of Quality and Reliability Management*, 22 (6):626–641, 2005.

C. Anderson-Cook. Evaluating the series or parallel structure assumption for system reliability. *Quality Engineering*, 21(1):88–95, 2008.

C. M. Anderson-Cook. Opportunities and issues in multiple data type meta-analyses. *Quality Engineering*, 21(3):241–253, 2009.

C. M. Anderson-Cook, T. Graves, M. S. Hamada, N. Hengartner, V. E. Johnson, C. S. Reese, and A. G. Wilson. Bayesian stockpile reliability methodology for complex systems with application to a munitions stockpile. *Journal of the Military Operations Research Society*, 12(2):25–38, 2007.

C. M. Anderson-Cook, T. Graves, N. Hengartner, R. Klamann, A. Koehler, A. G. Wilson, G. Anderson, and G. Lopez. Reliability modeling using both system test and quality assurance data. *Journal of the Military Operations Research Society*, 13:5–18, 2008.

T. Bedford, J. Quigley, and L. Walls. Expert elicitation for reliable system design. *Statistical Science*, 21(4):428–462, 2006.

J. O. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 3:385–402, 2006.

G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, MA, 1973.

G. S. Datta and M. Ghosh. On the invariance of noninformative priors. *The Annals of Statistics*, 24:141–159, 1996.

R. M. Dickinson, L. J. Freeman, B. A. Simpson, and A. G. Wilson. Statistical methods for combining information: Stryker family of vehicles reliability case study. *Journal of Quality Technology*, 47:400–415, 2015.

D. Gamerman and H. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Chapman & Hall/CRC, Boca Raton, FL, 2nd edition, 2006.

P. H. Garthwaite, J. B. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–700, 2005.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, FL, 3rd edition, 2013.

M. Ghosh. Objective priors: An introduction for frequentists with discussion. *Statistical Science*, 26(2):187–202, 2011.

M. Goldstein. Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, 3:403–420, 2006.

J. Guo and A. G. Wilson. Bayesian methods for estimating the reliability of complex systems using heterogeneous multilevel information. *Technometrics*, 55(4):461–472, 2013.

M. S. Hamada, A. G. Wilson, C. S. Reese, and H. F. Martz. *Bayesian Reliability*. Springer, New York, NY, 2008.

M. S. Hamada, A. G. Wilson, B. Weaver, R. Griffiths, and H. F. Martz. Bayesian binomial assurance tests for system reliability using component data. *Journal of Quality Technology*, 46(1):24–32, 2014.

J. G. Ibrahim and M.-H. Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

V. E. Johnson, T. Graves, M. S. Hamada, and C. S. Reese. A hierarchical model for estimating the reliability of complex systems. In J. Bernardo, M. Bayarri, J. Berger, A. David, D. Heckerman, A. Smith, and M. West, editors, *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, pages 199–213. Oxford University Press, Oxford, UK, 2003.

V. E. Johnson, A. Moosman, and P. Cotter. A hierarchical model for estimating the early reliability of complex systems. *IEEE Transactions on Reliability*, 54(2):224–231, 2005.

W. Q. Meeker and L. A. Escobar. Reliability: the other dimension of quality. *Quality Technology and Quantitative Management*, 1:125, 2004.

W. Q. Meeker and Y. Hong. Reliability meets big data: opportunities and challenges. *Quality Engineering*, 26(1):102–116, 2014.

M. Meyer and J. M. Booker. *Eliciting and Analyzing Expert Judgment.* ASA/SIAM, Philadelphia, PA, 2001.

M. Morgan and M. Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis.* Cambridge University Press, Cambridge, UK, 1991.

National Research Council Oversight Committee for the Workshop on Testing for Dynamic Acquisition of Defense Systems. In V. Nair and M. L. Cohen, editors, *Testing of Defense Systems in an Evolutionary Acquisition Environment.* National Academies Press, Washington, DC, 2006.

National Research Council Panel on Operational Test Design and Evaluation of the Interim Armored Vehicle. *Improved Operational Testing and Evaluation Methods of Combining Test Information for the Stryker Family of Vehicles and Related Army Systems: Phase II Report*. National Academies Press, Washington, DC, 2004.

National Research Council Panel on Reliability Growth Methods for Defense Systems. *Reliability Growth: Enhancing Defense System Reliability*. National Academies Press, Washington, DC, 2015.

National Research Council Panel on Statistical Methods for Testing and Evaluating Defense Systems. In M. L. Cohen, J. E. Rolph, and D. L. Steffey, editors, *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements*. National Academies Press, Washington, DC, 1998.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Chichester, UK, 2006.

J. Parker. Bayesian prior distributions for multi-component systems. *Naval Research Logistics Quarterly*, 19(3):509–515, 1972.

C. S. Reese, J. A. Calvin, J. C. George, and R. J. Tarpley. Estimation of fetal growth and gestation in bowhead whales. *Journal of the American Statistical Association*, 96(455):915–923, 2001.

C. S. Reese, A. G. Wilson, M. S. Hamada, H. F. Martz, and K. Ryan. Integrated analysis of computer and physical experiments. *Technometrics*, 46(2):153–164, 2004.

M. J. Rhoads. Design for reliability handbook. Technical Report TR-2011-24, U.S. Army Materiel Systems Analysis Activity, Aberdeen Proving Grounds, MD, 2011. URL `http://www.amsaa.army.mil/Documents/CRG/Design%20for%20Reliability%20Handbook%20(TR-2011-24).pdf`.

C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, NY, 2010.

U. S. Nuclear Regulatory Commission. A review of NRC staff uses of probabilistic risk assessment. Technical Report NUREG-1489, U. S. Nuclear Regulatory Commission, Washington, DC, 1994.

D. Von Winterfeldt and W. Edwards. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, UK, 1986.

A. G. Wilson and A. V. Huzurbazar. Bayesian networks for multilevel system reliability. *Reliability Engineering and Systems Safety*, 92(10):1413–1420, 2007.

A. G. Wilson, L. McNamara, and G. D. Wilson. Information integration for complex systems. *Reliability Engineering and System Safety*, 92:121–130, 2007.

X. Zhang and A. G. Wilson. System reliability and component importance under dependence: A copula approach. *Technometrics*, 2016.

R. Zoh. *Using the Negative Log-Gamma Distribution for Bayesian System Reliability Assessment*. PhD thesis, Iowa State University, Department of Statistics, 2012.