Bayesian vs frequentist techniques for the analysis of binary outcome data

By M. Stapleton

Abstract

We compare Bayesian and frequentist techniques for analysing binary outcome data. Such data are commonly found in defence applications, where the outcome can be encoded as one of two discrete values. Examples include detection (detected / not detected), armour or ammunition testing (penetrates / does not penetrate) and component testing (pass / fail). We apply both frequentist and Bayesian techniques to component testing data, where an impulse with specific energy is applied and a single pass/fail outcome is observed. The frequentist method allows us to estimate the probability of failure as a function of energy, as well as estimate confidence intervals. The Bayesian method similarly allows us to estimate the probability of failure and calculate confidence (credible) intervals, generating results that are comparable to the frequentist analysis. However, with both analyses we find ourselves considering probabilistic questions about the probability of failure. This layering of probabilities can make the results difficult to interpret and careless choice of significance level relative to the probability of failure can lead to conclusions that are statistically valid yet highly questionable. Only the Bayesian approach provides a method for combining these probabilities, through calculation of the predictive probability of failure: the probability that the next component will fail for a particular energy. This technique provides a more intuitive interpretation of the results and applying it to our component testing data we identify a weakness in the underlying statistical model that had not previously been spotted.

1. Introduction

Binary outcomes occur in many defence applications; usually when a system or component is being tested and a simple pass/fail result is recorded. Examples include detection (detected / not detected), armour or ammunition testing (penetrates / does not penetrate) and component testing (pass / fail). In many instances, we are interested in the probability of each outcome as some parameter is varied; the speed of the bullet in armour testing for instance, and would like to make inferences about the probability of pass/fail as that parameter is varied. This problem can be approached using either frequentist or Bayesian methods and in this paper we describe and compare the two. We show how to make inferences in each case and show that the Bayesian technique has a particularly useful device: the predictive probability of pass/fail, which provides a more intuitive interpretation of the results and highlights a potential weakness in our model that we were not previously aware of.

The remainder of this paper is structured as follows: in the next subsections we introduce data and a model which we shall use for the subsequent analysis. Sections 2 and 3 then demonstrate the use of confidence bands and intervals for estimation. Confidence bands are treated first as it turns out they are slightly easier mathematically. We discuss the results in section 4, which leads us to consider the predictive probability of failure in section 5. Final conclusions are given in section 6.

1.1. *Data and model*

To aid the discussion, we shall use as an example a data set from two component testing trials at AWE. An impulse with specific energy, E (in arbitrary units), is applied to a component and a single pass/fail outcome is observed. This was carried out for two types of components, A and B, during two separate trials and in each case we wish to infer the probability of failure as a function of energy. We shall model the underlying distribution using the cumulative lognormal,

$$\text{Prob}(\text{Failure}|E, \mu, \sigma) = \Phi\left(\frac{\log E - \mu}{\sigma}\right) \tag{1.1}$$

where Φ is the cdf for the standard normal distribution and μ and σ are unknown parameters.

Throughout the paper we will denote the true probability of failure as a function of energy as $\theta(E; \mu, \sigma)$, though sometimes we will drop the explicit μ and σ dependence: $\theta(E)$. We use this notation rather than $\text{Prob}(\text{Failure}|E, \mu, \sigma)$ as the latter can lead to confusion owing to the many 'probabilities' that arise in this

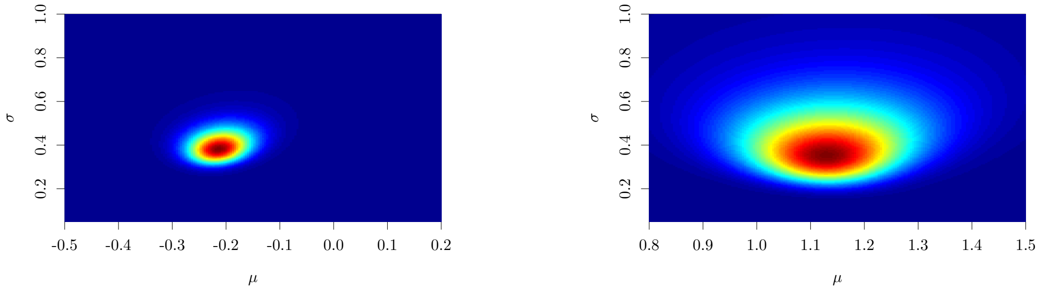

Figure 1. Pseudocolour plot of the likelihood function for component A (left) and component B (right). Red denotes high values and blue low values – the colour scales are normalised for each plot since the likelihood is a function of the number of data points which is different for the two trials, so the likelihoods are not directly comparable.

problem. Moreover, we emphasise that the quantity of interest is the probability of failure, $\theta(E)$, not μ and σ directly. This is important as, for any *given* E, there are infinitely many combinations of μ and σ that will give rise to the same $\theta(E)$, namely those satisfying $\log(E) - \mu = \text{const} \times \sigma$, and this will have consequences for the analysis.

The approach is therefore one of inferring properties of the probability of failure, $\theta(E)$, via the parameters μ and σ. Our focus will be on calculating confidence intervals† for the probability of failure as a function of energy. Generally speaking, these will be quantities $L(E)$ and $U(E)$ which, in a sense determined by whatever statistical principle we use, contain the true value with some probability. This can be done pointwise, or as confidence bands: the pointwise confidence interval $[L_p(E_0), U_p(E_0)]$ for a specific energy, E_0, expects the true probability of failure at that energy, $\theta(E_0)$, to be within the interval with a certain probability. The confidence band, $[L_b(E), U_b(E)]$, requires that the true curve for the probability of failure, $\theta(E)$, is within the interval for all values of E. For a precise distinction we need to specify whether we are using frequentist or Bayesian statistics, which we shall defer to the relevant sections.

1.2. *Likelihood function*

Both frequentist and Bayesian methods require a likelihood function, which is derived from the model, Eq (1.1). We consider each data point as the pair (E_i, X_i), where E_i is the energy of the ith test and X_i is 1 if the device failed and 0 if it did not. Hence, the likelihood is

$$\text{Prob}(X|\mu,\sigma) = \prod_i \text{Prob}(X_i|E_i,\mu,\sigma) \tag{1.2}$$

where we write X to denote all the data. From Eq (1.1),

$$\text{Prob}(X_i|E_i,\mu,\sigma) = \begin{cases} \Phi\left(\frac{\log E_i - \mu}{\sigma}\right) & X_i = 1 \\ 1 - \Phi\left(\frac{\log E_i - \mu}{\sigma}\right) & X_i = 0 \end{cases} \tag{1.3}$$

$$= X_i \Phi\left(\frac{\log E_i - \mu}{\sigma}\right) + (1 - X_i)\left[1 - \Phi\left(\frac{\log E_i - \mu}{\sigma}\right)\right]. \tag{1.4}$$

The likelihood as a function of μ and σ is plotted for both components in figure 1.

We shall denote maximum likelihood estimates with a hat, i.e. $\hat{\mu}$ and $\hat{\sigma}$. The maximum likelihood values for the model parameters obtained by the fit are $\hat{\mu} = -0.214$ and $\hat{\sigma} = 0.383$ for component A and $\hat{\mu} = 1.13$ and $\hat{\sigma} = 0.355$ for component B. Hence, on the face of the maximum likelihood estimate, it seems that component A fails at lower energies and has a slightly wider intermediate range, where the probability of failure is traversing between 0 and 1. This is illustrated in figure 2, which shows the data for each trial with their maximum likelihood fits, $\theta(E; \hat{\mu}, \hat{\sigma})$.

† Bayesian confidence intervals are sometimes called ‘credible’ intervals, though we shall not use this terminology.

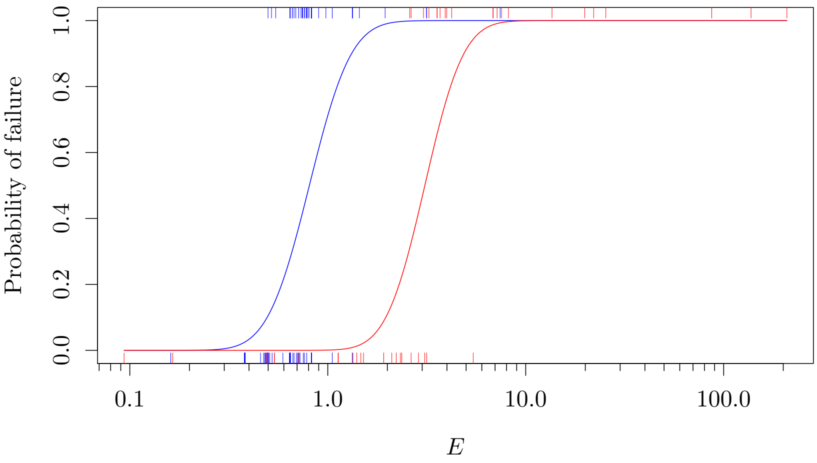

Figure 2. A plot of the maximum likelihood fits, $\theta(E;\hat{\mu},\hat{\sigma})$ for components A (blue) and B (red). The original data is also shown as rug plots: the points on the top of the plot correspond to failures ($X_i = 1$); the points on the bottom correspond to non-failure ($X_i = 0$).

2. Confidence bands

Generally speaking, a confidence band $[L_b(E), U_b(E)]$ satisfies

$$\text{Prob}\Big\{L_b(E) \le \theta(E) \le U_b(E),\ \forall E\Big\} = 1 - \alpha. \tag{2.1}$$

In other words, with probability $1-\alpha$ the true curve, $\theta(E)$, falls inside the interval $[L_b(E), U_b(E)]$ for all E. The difference between the frequentist and Bayesian approach to this problem is in what we consider to be the random variable in Eq (2.1). The frequentist regards $\theta(E)$ to be an unknown yet fixed quantity, in which case $[L_b(E), U_b(E)]$ needs to be a random interval constructed from the random variables (actually, functions) $L_b(E)$ and $U_b(E)$ that depend on the data, X. These must have probability $(1-\alpha)$ of containing the true $\theta(E;\mu,\sigma)$ for all E, *whatever* the true values μ and σ. Conversely, the Bayesian assigns the unknown $\theta(E)$ a probability distribution (the posterior probability) as though it were a random variable. The Bayesian then needs to find curves $l_b(E)$ and $u_b(E)$ such that $\theta(E)$ has probability $(1-\alpha)$ of being between the two.

Hence, the frequentist confidence band is the random variable $[L_b(E), U_b(E)]$ and the Bayesian confidence band is the interval $[l_b(E), u_b(E)]$. To highlight the difference, we have adopted the convention in statistics to write random variables in upper case and values in lower case.

2.1. *Frequentist confidence bands*

To approach this problem, suppose we have conducted a hypothesis test and derived a critical region of size α in (μ,σ) space, which we denote S_α. There are many interesting connections between the spaces (μ,σ) and $(E,\theta(E;\mu,\sigma))$ which are beyond the scope of this paper, but it is straightforward to show that there is a one-to-one correspondence between convex regions of (μ,σ) space and bands in (E,θ) space, i.e. for any convex region, C, there is a band $[L_c(E), U_c(E)]$ such that $\theta(E,\mu,\sigma) \in [L_c(E), U_c(E)]$ for all E if and only if $(\mu,\sigma) \in C$. Hence, if we take the convex hull of S_α, which we denote C_α, then C_α will contain the true (μ,σ) with probability at least $1-\alpha$ and the corresponding band $[L_c(E), U_c(E)]$ will contain the true $\theta(E)$ curve with probability $\ge 1-\alpha$. The fact that this is an *in*equality comes from the fact that even if the true (μ,σ) is rejected by the hypothesis test, and so falls outside of S_α, it is still possible for it to fall inside the convex hull C_α. In order to gain equality, we would need the critical region to be always convex. This might be true, though proving it is left for future work.

Since confidence bands can be constructed from a critical region in (μ,σ) space, all we need is a method for generating a critical region. This can be done approximately by means of the likelihood ratio test: we use the test statistic

$$D = -2\log\frac{L(\mu,\sigma;X)}{L(\hat{\mu},\hat{\sigma};X)} \tag{2.2}$$

which is asymptotically chi-square distributed with two degrees of freedom (Mood 1963). This allows us to define a critical region by rejecting those (μ, σ) for which the test statistic is larger than the $1 - \alpha$ quantile of χ^2_2. The confidence band is then formed by calculating

$$L_b(E) = \inf_{\mu,\sigma \in C_\alpha} \theta(E; \mu, \sigma) \tag{2.3}$$

$$U_b(E) = \sup_{\mu,\sigma \in C_\alpha} \theta(E; \mu, \sigma) \tag{2.4}$$

for all E.

2.2. *Bayesian confidence bands*

For a Bayesian analysis, we need to calculate the posterior probability density for the curve $\theta(E; \mu, \sigma)$ given the data. However, we shall find it more convenient to consider instead the joint posterior density over μ and σ:

$$p(\mu, \sigma | X) \propto \mathrm{Prob}(X | \mu, \sigma) p(\mu, \sigma) \tag{2.5}$$

where $p(\mu, \sigma)$ is the prior on μ and σ. The proportionality factor is simply a constant chosen so that $p(\mu, \sigma | X)$ is a normalised probability density, i.e. the integral of $p(\mu, \sigma | X)$ over all μ and σ is equal to one.

For the priors, we shall assume independent Jeffrey's priors (Jaynes 2003):

$$p(\mu, \sigma) = p(\mu) p(\sigma) \tag{2.6}$$

$$p(\mu) \propto 1, \quad \mu \in [\mu_0, \mu_1] \tag{2.7}$$

$$p(\sigma) \propto \frac{1}{\sigma}, \quad \sigma \in [\sigma_0, \sigma_1]. \tag{2.8}$$

The limits μ_0, μ_1, σ_0 and σ_1 are introduced as a technicality to enable numerical integration of the posterior; usually we can make the theoretical limits infinite, and though these lead to improper priors, they have no effect on the results†. We have chosen values considered reasonable for this problem: $\mu_0 = -1$, $\mu_1 = 3$, $\sigma_0 = 0.006$ and $\sigma_1 = 3$, though the precise values do not matter providing they define a sufficiently large region of (μ, σ) relative to the location of the mass of the likelihood. Calculations are carried out numerically on a grid of μ and σ values and the normalisation in Eq (2.5) is carried out by a basic numerical integration.

A Bayesian $(1 - \alpha)$ confidence band is a pair of curves, $l_b(E)$ and $u_b(E)$, such that the true curve $\theta(E)$ lies between the two for all E with probability $1 - \alpha$. We take advantage again of the correspondence between bands and convex regions in (μ, σ) space: for any convex region, C, there is a band $[l(E), u(E)]$ such that $\theta(E, \mu, \sigma) \in [l(E), u(E)]$ for all E if and only if $(\mu, \sigma) \in C$. Hence,

$$\mathrm{Prob}\left\{\theta(E, \mu, \sigma) \in [l(E), u(E)], \forall E | X\right\} = \mathrm{Prob}\left\{(\mu, \sigma) \in C | X\right\} \tag{2.9}$$

$$= \int_C p(\mu, \sigma | X) \; d\mu \; d\sigma \tag{2.10}$$

and a convex region, C_α defines a Bayesian $(1 - \alpha)$ confidence band if this probability is equal to $1 - \alpha$, i.e.

$$\int_{C_\alpha} p(\mu, \sigma | X) \; d\mu \; d\sigma = 1 - \alpha. \tag{2.11}$$

This relation does not uniquely determine C_α, so we shall add an additional constraint: that the posterior probability of all points inside C_α are larger than all points outside. This allows for a straightforward determination of C_α numerically from the calculated posterior distribution $p(\mu, \sigma | X)$ by sorting points by value and accumulating them until the sum equals $1 - \alpha$. As before, we don't know whether this will generally lead to a convex region, though for our data it does and we associate it with C_α and calculate the corresponding confidence intervals as before, using Eqs (2.3) and (2.4).

2.3. *Results*

The resulting confidence bands are shown in figures 3 and 4 for $\alpha = 0.05$. The frequentist and Bayesian curves, whilst being calculated using very different methods, generate almost indistinguishable results.

† This is not strictly true for the upper limit on σ since the likelihood does not go to zero as $\sigma \to \infty$. This means that the limit will begin to have a noticeable impact as σ_1 becomes very large (many orders of magnitude). However, since we can reasonably expect σ not to be anywhere near that big, we can set σ_1 to a reasonable value where the precise choice does not matter.

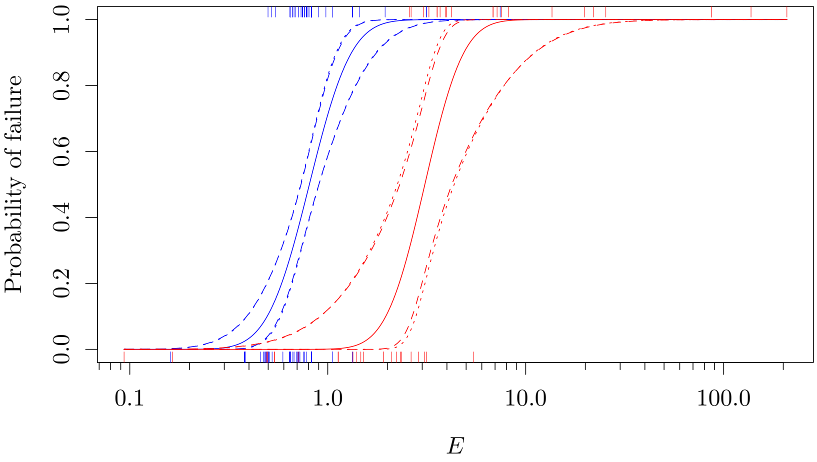

FIGURE 3. Plot of the 0.95 confidence bands for components A (blue) and B (red). Frequentist confidence bands are shown with dashed lines and Bayesian confidence bands with dotted lines. In this case, the curves are almost indistinguishable.

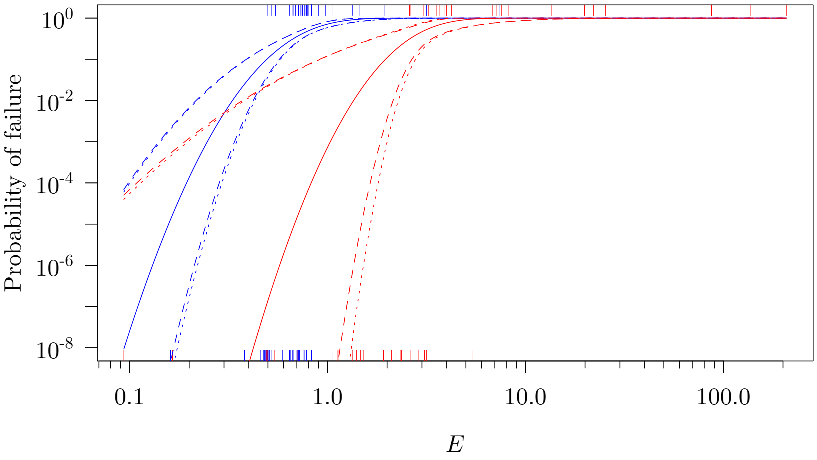

FIGURE 4. Plot of the 0.95 confidence bands for components A (blue) and B (red) on a logarithmic y scale. Frequentist confidence bands are shown with dashed lines and Bayesian confidence bands with dotted lines. In this case, the curves are almost indistinguishable.

3. Confidence intervals

We wish to find the pointwise confidence interval at specific energy E_0,

$$\text{Prob}\Big\{L_p(E_0) \leq \theta(E_0) \leq U_p(E_0)\Big\} = 1 - \alpha \tag{3.1}$$

Being more precise, as before, the frequentist wishes to find random variables $L_p(E_0)$ and $U_p(E_0)$ such that the interval $[L_p(E_0), U_p(E_0)]$ contains the true value, $\theta(E_0)$ with probability $1 - \alpha$ whatever the true value of $\theta(E_0)$. The Bayesian regards the posterior distribution on $\theta(E_0)$ and wishes to find the interval $[l_p(E_0), u_p(E_0)]$ such that $\theta(E_0)$ has probability of $1 - \alpha$ of being within this interval.

3.1. *Frequentist confidence intervals*

Our approach will be to use the generalised likelihood ratio test (Mood 1963): say we have a two-dimensional parameter that we wish to make inferences on, $\phi = (\phi_1, \phi_2)$, but only wish to test the hypothesis that $\phi_1 = a$

without specifying ϕ_2, then we can do this by comparing the test statistic,

$$\Lambda = -2\log\left[\sup_{\phi_2}\frac{L(a,\phi_2;X)}{L(\hat{\phi_1},\hat{\phi_2};X)}\right] \tag{3.2}$$

to the chi-square distribution with one degree of freedom. In our case, we wish to test the hypothesis that $\theta(E_0)$ takes some value, θ_0, regardless of μ and σ. Of course, given μ and σ, $\theta(E_0)$ is fully determined, so we need to re-parameterise in terms of $\theta(E_0)$ and one of the other parameters. The correct parameterisation in this case is $(\theta(E_0),\sigma)$, since for any combination of these we can uniquely determine μ:

$$\mu(\theta,\sigma) = \log E_0 - \Phi^{-1}(\theta)\sigma. \tag{3.3}$$

Note that the parameterisation in terms of $(\theta(E_0),\mu)$ does not work as σ is undetermined as a function of these parameters when $\theta(E_0) = 0.5$.

We can therefore test the hypothesis that $\theta(E_0) = \theta_0$ by comparing the statistic

$$\Lambda(\theta_0) = -2\log\left[\sup_{\sigma}\frac{L(\mu(\theta_0,\sigma),\sigma;X)}{L(\hat{\mu},\hat{\sigma};X)}\right] \tag{3.4}$$

to the $(1-\alpha)$ quantile of the chi-square distribution with one degree of freedom. The critical region for $\theta(E_0)$ is then precisely our confidence interval.

3.2. *Bayesian confidence intervals*

For a given energy, E_0, we wish to find the $(1-\alpha)$ confidence interval $[l_p(E_0),u_p(E_0)]$ such that the posterior probability that $\theta(E_0)\in[l_p(E_0),u_p(E_0)]$ is equal to $1-\alpha$. More specifically, we shall seek the symmetric confidence interval,

$$\text{Prob}\left\{\theta\le l_p|X\right\} = \frac{\alpha}{2} \tag{3.5}$$

$$\text{Prob}\left\{\theta\ge u_p|X\right\} = \frac{\alpha}{2}. \tag{3.6}$$

Now, the probability that $\theta\le l_p$ is trivially equal to the probability that

$$\Phi\left(\frac{\log E_0-\mu}{\sigma}\right)\le l_p. \tag{3.7}$$

In other words, for a given E_0, it is the probability that μ and σ take values that satisfy this relationship. Since Φ is monotonic increasing, we can rewrite this relationship as

$$\frac{\log E_0-\mu}{\sigma}\le\Phi^{-1}(l_p) \tag{3.8}$$

$$\log E_0\le\mu+\sigma\Phi^{-1}(l_p). \tag{3.9}$$

We therefore see that the probability of $\theta\le l_p$ is equal to the probability that $\mu+\sigma\Phi^{-1}(l_p)\ge\log E_0$. This is simply the integral of $p(\mu,\sigma|X)$ over that region,

$$\text{Prob}\left(\theta(E_0;\mu,\sigma)\le l_p|X\right) = \int_{-\infty}^{\infty}d\mu\int_0^{\infty}d\sigma\ H\left(\mu+\sigma\Phi^{-1}(l_p)-\log E_0\right)\ p(\mu,\sigma|X) \tag{3.10}$$

where $H(x)=0$ for $x<0$ and $H(x)=1$ for $x\ge0$. We then find $l_p(E_0)$ such that this integral is equal to $\alpha/2$, which is straightforward to do numerically. An almost identical argument holds for $u_p(E_0)$.

3.3. *Results*

Pointwise confidence intervals for $\alpha = 0.05$ are shown in figure 5 and 6. There is more difference between the Bayesian and frequentist confidence intervals than we observed for the confidence bands, though they are nonetheless still quite similar.

4. Discussion

In the previous sections we have demonstrated how to calculate frequentist and Bayesian confidence bands and intervals. Despite the fundamentally different philosophical foundations and methods of calculation, the results have come out quite similar. This is perhaps not as surprising as it may seem for two important reasons:

First, the Bayesian analysis assumed non-informative priors, which are well known to bring Bayesian re-

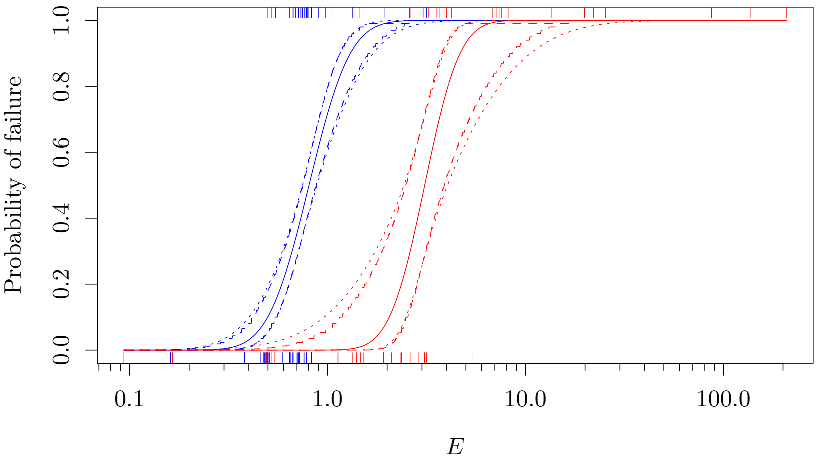

Figure 5. Plot of the 0.95 pointwise confidence intervals for components A (blue) and B (red). Frequentist confidence intervals are shown with dashed lines and Bayesian confidence intervals with dotted lines.

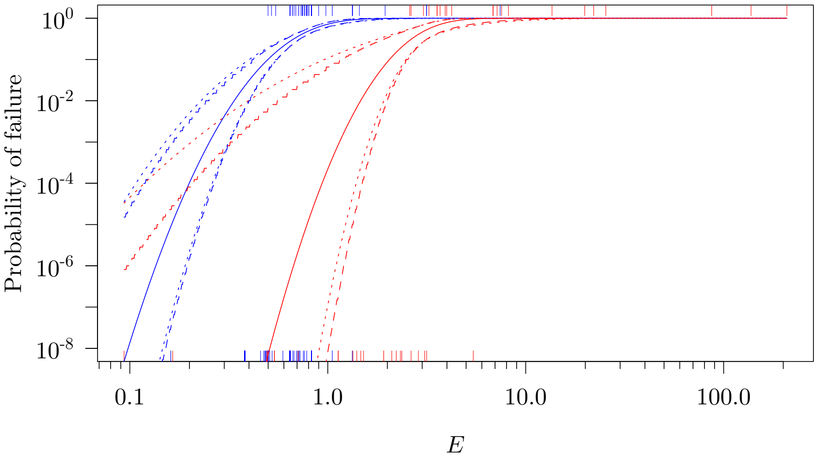

Figure 6. Plot of the 0.95 pointwise confidence intervals for components A (blue) and B (red) on a logarithmic y scale. Frequentist confidence intervals are shown with dashed lines and Bayesian confidence intervals with dotted lines.

sults more in line with the corresponding frequentist ones (Jaynes 2003). Bayesian confidence intervals using informative priors may differ significantly from their frequentist counterparts.

Second, in many respects the Bayesian analysis is the Bayesian version of an essentially frequentist procedure: we have some parameter that we are trying to estimate from data and express uncertainty in our estimate through the use of confidence intervals. The parameter we are estimating in this case is the probability of failure as a function of E, and we express this parametrically through a model (Eq (1.1)) with parameters μ and σ. We can then analyse the uncertainty in the probability of failure, generating confidence intervals as we have already shown. However, there are disadvantages to this, such as having to choose a value for α. For instance, we have generated confidence intervals for $\alpha = 0.05$, which will contain the true value with probability 0.95 (in some sense). This may be reasonable if we are interested in regions of energy where the probability of failure is commensurate with α, i.e. around 0.05. However, in many safety-critical applications we require much lower probabilities and it is not immediately clear what confidence level we should choose for estimating the probability of failure when it is very small (e.g. $\sim 10^{-6}$). For instance, we might decide on a confidence level of $1 - \alpha = 0.99$, but there is a contradiction if we are saying that we need a probability of failure to be as small as 10^{-6}, yet we are happy to be wrong in this estimate with probability 10^{-2}. One might therefore take a more extreme α of 10^{-6}, but this will result in extremely wide confidence intervals. The problem is that confidence intervals handle uncertainty by needing to be 'correct' with a certain probability; they do not take into account how wrong our estimate might be when it does go wrong.

The Bayesian method offers a route out of this problem with the *predictive probability of failure*: combining all the probabilities we have into a final probability of failure, which is our subjective probability, given our data and model, that the next component will fail if subjected to energy E. We do this by means of a 'predictive' calculation (Aitchison 1975): we incorporate all the uncertainty in our model (plus assumptions) to provide an overall probability statement predicting the outcome of the next experiment. It is a purely Bayesian concept, with no frequentist analogue†.

5. Predictive probability of failure

The parameter $\theta(E)$ represents the probability of failure when an impulse of energy E is applied to the component. This is a physical property of the component and if we could test a very large number of them it would represent the fraction that failed under an impulse energy, E. In principle, it could also be calculated from fundamental physics. However, in reality we do not know the value for θ and all we have is data that, along with a model, allows us to quantify our lack of knowledge. Using the Bayesian framework, our data allows us to determine a (posterior) probability distribution over θ, which we denote $p(\theta|E, X)$, and this is a measure of our belief for θ taking specific values at some energy E. For instance $p(\theta = 0.1|E = 1, X)$ is a measure of our belief that θ takes the value 0.1 when $E = 1$. By gathering more data, we can make the distribution $p(\theta|E, X)$ tighter, which amounts to narrowing down the range of θ that we believe is true. However, we will always have some residual uncertainty over its actual value.

We now wish to ask: given everything we know so far, what is the probability that the next component will fail given an impulse of energy E? The answer is subjective and depends on all our knowledge up to this point. We call this the 'predictive probability', Prob(Failure$|E, X$), and it reflects our belief *at the time of carrying out the experiment* that the next component will fail at energy E. It is not the outcome of a long-run series of experiments, and cannot be derived from fundamental physics. To calculate the predictive probability, we reason in the following way: suppose that there are only a small set of values that θ could take, i.e. θ_1, θ_2, $\ldots\theta_N$. The probability that the next component will fail is equal to: the probability that θ_1 is the true value, $p(\theta_1|E, X)$, multiplied by the probability of failure in this case, θ_1; plus the probability that θ_2 is the true value, $p(\theta_2|E, X)$, multiplied by the probability of failure in this case, θ_2; etc. In other words, it is the 'average' probability that the components will fail. Extending this argument to a continuum of values for θ, we have

$$\text{Prob(Failure}|E, X) = \int_0^1 \theta \; p(\theta|E, X) d\theta \tag{5.1}$$

which is the mean of the posterior density $p(\theta|E, X)$.

Another way we could write Eq (5.1) is to note that $\theta = \text{Prob(Failure}|E, \theta)$, in which case

$$\text{Prob(Failure}|E, X) = \int_0^1 d\theta \; \text{Prob(Failure}|E, \theta) \; p(\theta|E, X). \tag{5.2}$$

This is simply the decomposition of a probability in terms of θ, which we can do regardless of what θ means. Hence, instead of decomposing the probability in terms of θ, we can also do it in terms of μ and θ, in which case we have

$$\begin{aligned} \text{Prob(Failure}|E, X) &= \int_0^\infty d\sigma \int_{-\infty}^\infty d\mu \; \text{Prob(Failure}|E, \mu, \sigma) p(\mu, \sigma|X) & (5.3) \\ &= \int_0^\infty d\sigma \int_{-\infty}^\infty d\mu \; \Phi\left(\frac{\log E - \mu}{\sigma}\right) p(\mu, \sigma|X). & (5.4) \end{aligned}$$

The right hand side of Eq (5.4) is then straightforward to compute numerically.

We show the results for Prob(failure$|E, X$) in figure 7, with low energies shown in figure 8. The predictive probabilities are plotted as well as frequentist confidence intervals with $\alpha = 10^{-6}$, which we might have used for estimating probabilities at the 10^{-6} level. It is clear from the plot that the predictive probability provides a much less pessimistic result than the confidence interval since it explicitly takes into account all of the possible values for $\theta(E)$.

The results for the predictive probability also highlight an interesting feature of this model, which was not apparent in previous results. The predictive probability of failure for component A falls below that of component

† Frequentist prediction intervals do exist, but they cannot be applied to binary outcome data.

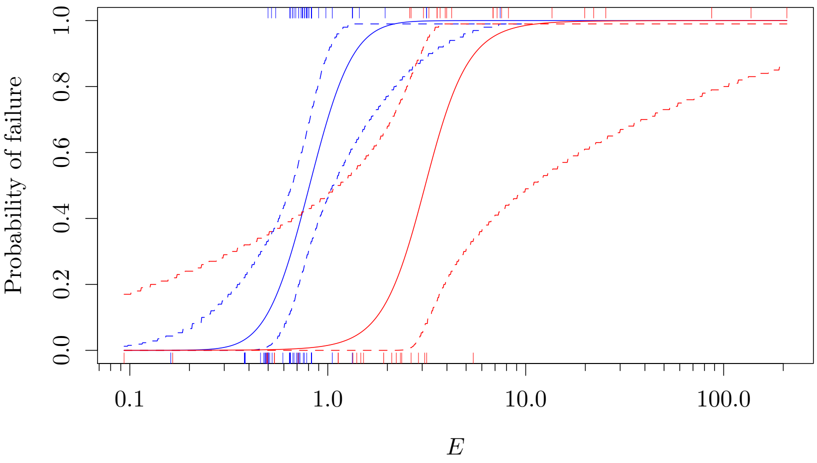

FIGURE 7. Predictive probability of failure for components A (blue) and B (red). The solid curves show the predictive probability of failure. The dashed curves show the confidence interval using $\alpha = 10^{-6}$.

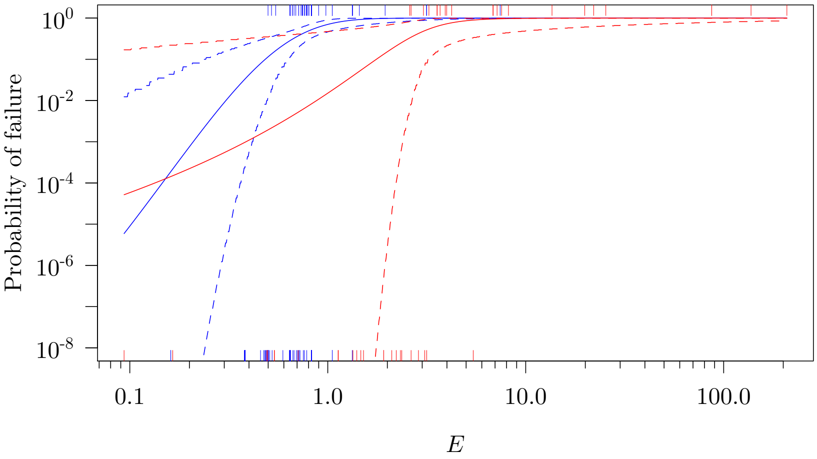

FIGURE 8. Predictive probability of failure for components A (blue) and B (red) on a logarithmic y scale. The solid curves show the predictive probability of failure. The dashed curves show the confidence interval using $\alpha = 10^{-6}$.

B for energies below around $E = 0.2$. This means that, below about $E = 0.2$ we expect component B to be more likely to fail than component A, but this is in contradiction with the fact that all of the component B failures occurred at higher energies than those for component A. i.e. taking the data at face value, we would expect to require a higher energy impulse to make component B fail, yet the predictive probability suggests that at low energies component B is more likely to fail than component A.

Mathematically, the higher predictive probability of failure for component B is caused by the relatively wide range of σ that are compatible with the data, combined with a likelihood function that is symmetric about $\log E = \mu$: large σ causes the curve to decay slowly for both small and large energies, so if there is a wide range of relatively large σ that is supported by the data this will generate a predictive probability that also decays slowly. Hence, our predictive probability is being heavily influenced by our choice of a symmetric likelihood function so any conclusions we draw from the analysis should take into account how much we trust our model. This is an important observation, particularly since we are looking at a regime of energy outside most of the trials data, so our results will be heavily influenced by our choice of model.

6. Conclusions

We have demonstrated the use of frequentist and Bayesian methods for calculating various confidence intervals for the probability of failure of two different components. Further work could be carried out to formally compare the results for the two types of component, for instance by carrying out hypothesis testing that they have different model parameters. Again, there are frequentist and Bayesian versions for all of these analysis.

The frequentist and Bayesian confidence bands/intervals we calculated were all largely similar, though this is not necessarily a surprise as we have used non-informative priors for the Bayesian analysis. However, it is important to stress that the calculations are based on fundamentally different philosophies that are not equivalent. This is seen most noticeably when we ask the question 'how good are our inferences'? Frequentist results are designed to provide particular errors at a pre-specified rate, which can be verified by repeated experiment. For instance, we can assess the accuracy of the frequentist confidence intervals by repeatedly sampling (i.e. simulating) data from a known $\theta(E)$ and measuring the frequency with which the confidence intervals contain the true probability of failure. If the frequentist calculation is working well then this should be close to $1-\alpha$. Whilst such a procedure could be carried out for the Bayesian analysis, there is no fundamental reason why we should expect errors to occur at any particular rate except if, during our simulations, we also sample $\theta(E)$ from our prior distribution. Moreover, it slightly misses the point of the Bayesian analysis, which is to provide a logically consistent approach to dealing with uncertainty: to check that our Bayesian analysis is correct, we should merely have to check our logic (Jaynes 2003).

In practice, however, robustness is usually just as important for the statistician as accuracy. The statistician knows that whatever model they have chosen is probably 'wrong', so they want to have confidence that the techniques they use will generate reasonable answers despite the inadequacies in the model. This problem becomes even more acute when attempting to use the model to extrapolate beyond the bulk of the available data, as we attempted in the last section, and judicious analysis of the sensitivity of results to the model's assumptions is necessary in order to make conclusions from any analysis.

REFERENCES

A. M. Mood, F. A. Graybill and D. C. Boes *Introduction to the theory of statistics.* McGraw-Hill, 1963.

J. Aitchison and I. R. Dunsmore *Statistical Prediction Analysis.* Cambridge University Press, 1975.

E. T. Jaynes *Probability Theory The Logic of Science.* Cambridge University Press, 2003.