A Bayesian Model that Combines Disparate Evidence for the Quantitative Assessment of System Dependability

Bev Littlewood
David Wright
Centre for Software Reliability, City University,
Northampton Square, London EC1V 0HB, ENGLAND

Abstract

For safety-critical systems, the required reliability (or safety) is often extremely high. Assessing the system, to gain confidence that the requirement has been achieved, is correspondingly hard, particularly when the system depends critically upon extensive software. In practice, such an assessment is often carried out rather informally, taking account of many different types of evidence—experience of previous, similar systems; evidence of the efficacy of the development process; testing; expert judgement, etc. Ideally, the assessment would allow all such evidence to be combined into a final numerical measure of reliability in a scientifically rigorous way. In this paper we address one part of this problem: we present a means whereby our confidence in a new product can be augmented beyond what we would believe merely from testing that product, by using evidence of the high dependability in operation of previous products. We present some illustrative numerical results that seem to suggest that such experience of previous products, even when these have shown very high dependability in operational use, can improve our confidence in a new product only modestly.

1 Introduction

Critical systems are coming to depend more and more upon the correct functioning of software to ensure their safe operation. At the same time, the size and complexity of these software subsystems is increasing as designers take advantage of the extensive functionality that software makes possible—functionality that sometimes enhances different aspects of safety.

There are important unresolved questions concerning how one might go about designing such systems so that they will be sufficiently safe in operation. In this paper, however, we shall concentrate upon the difficult problems of *evaluation* that they pose. In particular, we shall be concerned with the problem of how to measure the reliability of such a software system when that reliability is likely to be very high.

In several recent papers different authors have pointed out some of the basic difficulties here, [Butler, Finelli 1993, Littlewood, Strigini 1993]. They

show that, if we are only going to use the evidence obtained from operational testing of the software, we shall only be able to make quite modest claims for its reliability. For example, Littlewood and Strigini show that even in the most favourable situation of all, that of a system that has not failed during x hours of statistically representative operational testing, we can draw only the weak conclusion that there is a 50:50 chance that it will survive failure-free for the same time x in the future.

The limitations here seem intrinsic: they arise from the relative paucity of evidence (when compared with the stringency of the reliability level that needs to be demonstrated) and will not be ameliorated significantly by better statistical models. To make a very strong claim—that a particular system is ultra-reliable—needs a great deal of evidence. If that evidence comprises only observation of failure-free behaviour, then the length of time over which such behaviour is observed needs to be very great. To assure the reliability goals of certain proposed and existing systems, for example the 10^{-9} probability of failure per hour for the 'fly-by-wire' computer systems in civil aircraft [Rouquet, Traverse 1986], would clearly require the systems to be observed *and show no failures* for lengths of time that are many orders of magnitude greater than is practicable.

Faced with these limitations to what can be claimed from merely observing the system in operation, it has been suggested that we should instead base our evaluations upon *all* the disparate kinds of evidence that are available. These include, in addition to the operational data discussed above, evidence of the efficacy of the development methods utilised, experience in building similar systems in the past, competence of the development team, architectural details of the design, etc. Most of these other sources of evidence about the dependability of a system will involve a certain amount of engineering judgement in the evaluator, which might itself introduce further uncertainty and potentiality for error. In addition, there are serious unresolved difficulties in *combining* such disparate evidence in order to make a single evaluation of the overall dependability and thus to make a judgement of acceptability.

In this paper we shall consider only a small part of this problem. We shall treat in detail the situation where we wish to augment the evidence that can be gained from the operational testing of a particular product, by also taking into account the success (or not) in building 'similar' products in the past. An important special case, of course, is that where there is unreserved good news from these previous products—i.e. none of them has failed during operational use up till the present time.

It should be emphasised that the goal in all this work is to obtain a *quantification* of the reliability of a product. The model that is proposed in the next sections, therefore, requires us to make certain assumptions about the failure process, and about how we represent our beliefs about certain model parameters. We acknowledge that these assumptions can be questioned, and are certainly very difficult to validate. However, we believe that they are reasonably plausible. More importantly, our main aim is to demonstrate that this kind of evidence can only improve our confidence in the reliability of a product

quite modestly. Thus, we would regard a critique of our results on the grounds that they are not sufficiently conservative as being in the spirit of our own aims; suggestions, on the other hand, that the assumptions here can be modified in order to arrive at much higher confidence in product reliability we would regard with suspicion. It seems to us that, particularly in the case of safety-critical applications, it is safest to adopt a conservative view of the informativeness of evidence unless there are scientifically valid reasons to believe the contrary.

2 Modelling Approach

When we use evidence we have obtained from building and operating previous products in order to try to improve the accuracy of the predictions that we can make about the reliability of a novel product, we must take account of two kinds of uncertainty. In the first place, there will be uncertainty concerning the actual reliabilities that have been achieved by these earlier products. Even in those cases where there is extensive operating experience, we shall never know the true reliability of a product and will have to use an estimate based upon the data collected during its operation. In those situations where we are dealing with products that are likely to be very reliable, we shall probably only see a small number (or even none at all) of failures even in quite extensive periods of operation.

The second source of uncertainty will concern the 'similarity' of the products that have been observed in the past, and the 'similarity' of the one under study to these past products. In what follows, we shall assume that the probabilities of failure of the different products, past and present, can be assumed to be realisations of independent and identically distributed random variables. This assumption, although an idealisation, captures the essentials of what we mean by 'similarity'. Thus, it means that the actual reliabilities of the different products will be different, as is clearly the case in reality. We would not expect the reliabilities of, say, two versions of a software-based telephone switch to be identical, even though we might be prepared to agree that the problems posed, and the quality of the processes deployed in their solution, were similar. The notion of 'similarity' in the eye of an observer here seems to be equivalent to a kind of 'indifference'. You might agree that two different products were similar for the purposes of the current exercise if you were indifferent between them in reliability terms: if you were asked to predict which would be the most reliable, before seeing them in operation, you would have no preference. This is represented by their probabilities of failure being identically distributed random variables: any probability statements you would make about the reliabilities of products $\mathcal{A}$ and $\mathcal{B}$ will be identical. The important point here is that this interpretation of 'similarity' in terms of indifference does not mean that you believe that the two products will have identical reliabilities [Laprie 1992] - indeed you will know that the actual reliabilities of the products will differ.

The two sources of uncertainty here are both important. However, it is the nature of the uncertainty concerning 'how similar' the products actually are

that will be most difficult to estimate in practice, since this requires us to see as many different products as possible. It is far more likely that we have large quantities of information about a few products, than that we have information on many products.

Consider first the failure process of a *single* software product $\mathcal{A}$. Assume a Bernoulli trials process model of the failures of this product in a sequence of 'demands' with neither debugging, maintenance, nor significant variation in the 'stressfulness' of the software's operational environment. Thus, in the first n trials of product $\mathcal{A}$, let R be the random number of failures occurring and p be the probability of failure on demand. Then the distribution of R for fixed n and p is

$$R \,|\, n, p \sim \binom{n}{r} p^r \,(1-p)^{n-r} \tag{1}$$

Now think of p as unknown and construct a Bayesian model by assuming that p is a realisation of a random variable P having a parametric distribution

$$P \,|\, \theta \sim f_p(p|\theta)$$

with parameter θ. Here we can think of this distribution for P as representing the general reliability of products in a particular *product family*, perhaps produced by a single development team, using a common development method, and for similar applications. For example, a family of products known to have highly variable reliability levels would correspond to a distribution $f_p(p|\theta)$ with a large variance, whereas for another product family a high 'average' product reliability would correspond to a small mean for $f_p(p|\theta)$. If we fully understood the true variation in reliabilities of the products in each of these two product families then we could describe the two families by specifying two different P-distributions having the required characteristics and index these P-distributions with two different θ-values, θ_1 and θ_2, say. More generally, our parameter space $\mathcal{S}$, say, for θ, could be said to represent a set of different conceivable reliability characteristics each of which potentially characterises a different *family of similar products*. I.e., given sufficient data on the reliability variation amongst the products of a particular family, a value of θ (and hence a particular distribution $f_p(p|\theta)$) could in principle be assigned as descriptive of that variation. In this way, we have defined a model in which θ can be thought of as a product-family-characterising parameter. For a product chosen at random from those of a particular family of similar products (i.e. particular θ) and observed for a sequence of n demands, it follows that (R, P) has joint distribution

$$(R, P) \,|\, n, \theta \sim \binom{n}{r} p^r \,(1-p)^{n-r} \, f_p(p|\theta), \tag{2}$$

given n and θ. Integrating (2) over p gives the conditional distribution of R given n and θ as

$$R|n, \theta \sim \binom{n}{r} \int_0^1 p^r \,(1-p)^{n-r} \, f_p(p|\theta) \, dp \tag{3}$$

or, expressed in terms of moments of $f_p(\cdot|\theta)$,

$$R\,|\,n,\theta \sim \binom{n}{r}\mathbf{E}\big(P^r(1-P)^{n-r}|\theta\big)\ . \tag{4}$$

If we *observe* that $R = r$ failures actually occur during n demands, then we can condition on this data by normalising (2) to give the updated distribution

$$P\,|\,r,n,\theta \sim \frac{p^r\,(1-p)^{n-r}\,f_p(p|\theta)}{\int_0^1 p^r\,(1-p)^{n-r}\,f_p(p|\theta)\,dp} \tag{5}$$

of the probability of failure on demand for this program, given θ, n and the observation r.

The last three equations describe properties of a general mixture of Bernoulli trials processes, where $f_p(\cdot|\theta)$ is the mixing distribution. Note that although exchangeability[1] of the original Bernoulli trials process has not been lost by mixing the processes, the property that non-intersecting sections of the process are independently distributed does not hold in general for the resulting mixed process. In fact the number R' of failures in a subsequent set of n' demands on the same product now has an updated distribution obtainable from (5) as

$$\begin{aligned} R'\,|\,r,n,n',\theta &\sim \binom{n'}{r'}\frac{\int_0^1 p^{r+r'}\,(1-p)^{n+n'-r-r'}\,f_p(p|\theta)\,dp}{\int_0^1 p^r\,(1-p)^{n-r}\,f_p(p|\theta)\,dp}\,, \\ &= \binom{n'}{r'}\frac{\mathbf{E}\Big(P^{r+r'}(1-P)^{n+n'-r-r'}\,|\,\theta\Big)}{\mathbf{E}(P^r(1-P)^{n-r}\,|\,\theta)} \end{aligned} \tag{6}$$

given n, r.

The distributions which we have considered up till this point are parameterised by θ. We now adopt a Bayesian approach to handling this parameterisation by supposing a prior distribution

$$\Theta \sim \mathrm{Prior}_\theta(\theta)\,,$$

with support set $\theta \in \mathcal{S}$. If we plan to observe and predict reliability only of a single software product, this extension adds very little to the model as so far described, since, by integrating over θ, the model is reduced to a degenerate ($|\mathcal{S}| = 1$) case of the assumptions described earlier. (Simply replace $f_p(p|\theta)$ by $\int_{\theta\in\mathcal{S}} f_p(p|\theta)\mathrm{Prior}_\theta(\theta)\,d\theta$ in the distributions above.)

The idea of a prior distribution for θ becomes a useful concept, however, if we wish to address the problem of *learning* about a *distribution* of product reliabilities by observing the failure behaviour of *multiple* software products from a single family $\langle\mathcal{A}_i\rangle$, say, of similar products. We can then represent a

[1] i.e., the property that any permutation of a portion of the binary success-failure sequence has the same probability as the unpermuted sequence. Equivalently, we can say that the probability of a precise sequence of successes and failures during a specified interval of discrete time (say from the 10^{th} to the 20^{th} demand, inclusive) can be expressed as a function of the *number*, only, of successes during that interval.

conservative[2] version of a *process* concept for the trend of their reliabilities, from one product to the next, by modelling these products' individual failure processes as above with *different* p_i, and an assumption that each of these p_i arises *independently given* θ for some *unknown, common* parameter value θ characterising the entire family of products.

Thus θ and p now play distinct roles in terms of the model concepts: Whereas each p_i still captures a property of a single software product, θ now represents a common unknown characteristic of the whole family of similar products. To obtain the value of θ would be to capture the reliability-relevant characteristic which these software products all have in common. For this *multi*-product model, there is now a real purpose behind including separate distributional assumptions for firstly θ, and secondly p_i given θ. In the following, we do not in fact assume that θ can ever be known[3]. However, we assume that we hold *probabilistic prior beliefs about* θ (i.e. beliefs about the possible distributions $f_p(\cdot|\theta)$ of reliabilities of products belonging to the family $\langle \mathcal{A}_i \rangle$). Then, any observation of failure behaviour of any subset of the sequence $\langle \mathcal{A}_i \rangle$ can be regarded as information about θ which we will use in order to learn about θ by the usual Bayesian learning mechanisms. Thus the second stage of our doubly stochastic model is to represent our prior beliefs about a subjective random variable Θ of which the true value θ for our particular product family is a single unknown realisation.

Observe now that, conditionally given θ and $\langle n_i \rangle_{i=1}^k$, our independence assumption for the $\langle P_i \rangle$ tell us that the first k terms of our $\langle R_i \rangle$ sequence are jointly distributed

$$\langle R_i \rangle_{i=1}^k \big| (\langle n_i \rangle_{i=1}^k, \theta) \sim \prod_{i=1}^k \binom{n_i}{r_i} \int_0^1 p^{r_i} (1-p)^{n_i - r_i} f_p(p|\theta)\, dp\,. \tag{7}$$

Once we have executed these k software products and observed their failure behaviour (i.e., r_i failures out of n_i trials for each product $\mathcal{A}_i$) then we can regard (7) as the likelihood function $L\big(\theta; \langle n_i, r_i \rangle_{i=1}^k\big)$ of the parameter θ given this failure data. $L\big(\theta; \langle n_i, r_i \rangle_{i=1}^k\big)$ is a product involving combinatorial terms together with moments of the parametric distribution $f_p(\cdot|\theta)$

$$\langle R_i \rangle_{i=1}^k \big| (\langle n_i \rangle_{i=1}^k, \theta) \sim \prod_{i=1}^k \binom{n_i}{r_i} \mathbf{E}\big(P^{r_i}(1-P)^{n_i - r_i} \,|\, \theta\big)\ . \tag{8}$$

In §3 we make use of the factor of this likelihood which depends on θ,

$$L_k(\theta) \quad = \quad \prod_{i=1}^k \mathbf{E}\big(P^{r_i}(1-P)^{n_i - r_i} \,|\, \theta\big)$$

[2] in the sense that we desist from making any stronger assumption of any kind of systematic development of reliability from one product to the next. For example, we do not assume an increasing trend in reliabilities of different products in the family.

[3] Loosely, we can say that in order to *know* the value of θ characterising a family $\langle \mathcal{A}_i \rangle$ of products, we would require a very large amount of operational failure data on *each* of a very large number of products belonging to that family.—So that we could accurately describe from empirical data the shape of the distribution $f_p(\,\cdot\,|\theta)$.

$$= \prod_{i=1}^{k} \int_0^1 p^{r_i} (1-p)^{n_i - r_i} f_p(p|\theta)\, dp \tag{9}$$

3 Bayesian Updating of Distributions in the General Case

To implement the Bayesian learning about Θ given observation of $\langle r_i \rangle_{r=1}^{k}$ we would like to calculate the posterior distribution of Θ. Recalling that the prior for Θ is denoted Prior_θ, for θ lying in $\mathcal{S}$, then the required posterior distribution is proportional to the product of the prior distribution for Θ and the likelihood function evaluated as (7)

$$\Theta \,\big|\, \langle n_i, r_i \rangle_{i=1}^{k} \sim c L_k(\theta)\, \text{Prior}_\theta(\theta)$$

where c is a function of $\langle r_i, n_i \rangle$ not involving θ, i.e.

$$\Theta \,\big|\, \langle n_i, r_i \rangle_{i=1}^{k} \sim \frac{\left[\prod_{i=1}^{k} \int_0^1 p^{r_i} (1-p)^{n_i - r_i} f_p(p|\theta)\, dp\right] \text{Prior}_\theta(\theta)}{\int_{\theta \in \mathcal{S}} \left[\prod_{i=1}^{k} \int_0^1 p^{r_i} (1-p)^{n_i - r_i} f_p(p|\theta)\, dp\right] \text{Prior}_\theta(\theta)\, d\theta} \tag{10}$$

Equation (10) moves the focus of attention away from failure probabilities P_i of products $\mathcal{A}_i$ by the integrations over p. It is now of great interest to know an up-to-date distribution for P given what has been observed (in order to make predictions about a particular new product, for example). Then our learning could be expressed directly in terms of the changing nature of the current uncertainty about a failure probability of some particular product. At this stage it is instructive to distinguish between four different stages in our learning about one of the failure probabilities, say P_k. The first of these is the prior marginal distribution of P_k

$$P_k \sim \int_{\theta \in \mathcal{S}} f_p(p_k|\theta) \text{Prior}_\theta(\theta)\, d\theta\,, \tag{11}$$

which represents our initial state of uncertainty concerning the reliability of any given product, $\mathcal{A}_k$, prior to any observation either of that or of any other product's behaviour.

The second most trivial case—observing only the past failure behaviour of the specific product of interest—has effectively already been covered by (5). Substituting $\int_{\theta \in \mathcal{S}} f_p(p|\theta) \text{Prior}_\theta(\theta)\, d\theta$ for $f_p(p|\theta)$ in (5) gives a conditional distribution

$$P_k \,|\, n_k, r_k \sim \frac{p_k^{r_k} (1-p_k)^{n_k - r_k} \int_{\theta \in \mathcal{S}} f_p(p_k|\theta) \text{Prior}_\theta(\theta)\, d\theta}{\int_{\theta \in \mathcal{S}} \int_0^1 p^{r_k} (1-p)^{n_k - r_k}\, f_p(p|\theta)\, dp\, \text{Prior}_\theta(\theta)\, d\theta} \tag{12}$$

for P_k given n_k and r_k.

Thirdly, replacing k by $k-1$ in (10) and then substituting this distribution in place of $\mathrm{Prior}_\theta(\theta)$ in (11) (or, alternatively, directly substituting $n_k = r_k = 0$ in (14) below) gives the distribution

$$P_k \,\big|\, \langle n_i, r_i\rangle_{i=1}^{k-1} \sim \frac{\displaystyle\int_{\theta\in\mathcal{S}} f_p(p_k|\theta) \left[\prod_{i=1}^{k-1} \int_0^1 p^{r_i}\,(1-p)^{n_i-r_i}\; f_p(p|\theta)\,dp\right] \mathrm{Prior}_\theta(\theta)\,d\theta}{\displaystyle\int_{\theta\in\mathcal{S}} \left[\prod_{i=1}^{k-1} \int_0^1 p^{r_i}\,(1-p)^{n_i-r_i}\, f_p(p|\theta)\,dp\right] \mathrm{Prior}_\theta(\theta)\,d\theta} \tag{13}$$

of P_k given observation of the failure behaviour $\langle n_i, r_i\rangle_{i=1}^{k-1}$ *only* of *other* products $\langle \mathcal{A}_i\rangle_{i=1}^{k-1}$.

Finally, replacing k by $k-1$ in (10) and then substituting this distribution in place of $\mathrm{Prior}_\theta(\theta)$ in (12) gives the distribution

$$P_k \,\big|\, \langle n_i, r_i\rangle_{i=1}^{k} \sim \frac{p_k^{r_k}\,(1-p_k)^{n_k-r_k} \displaystyle\int_{\theta\in\mathcal{S}} f_p(p_k|\theta) \left[\prod_{i=1}^{k-1} \int_0^1 p^{r_i}\,(1-p)^{n_i-r_i}\; f_p(p|\theta)\,dp\right] \mathrm{Prior}_\theta(\theta)\,d\theta}{\displaystyle\int_{\theta\in\mathcal{S}} \left[\prod_{i=1}^{k} \int_0^1 p^{r_i}\,(1-p)^{n_i-r_i}\, f_p(p|\theta)\,dp\right] \mathrm{Prior}_\theta(\theta)\,d\theta} \tag{14}$$

for P_k given observation both of the failure behaviour $\langle n_k, r_k\rangle$ of the product $\mathcal{A}_k$ itself and *also* the failures $\langle n_i, r_i\rangle_{i=1}^{k-1}$ of other products $\langle \mathcal{A}_i\rangle_{i=1}^{k-1}$.

4 The No-Failures Case

Consider the special case in which no failures at all have been observed—neither of the product for which we wish to predict reliability, nor of other products within the same product family. This case is of particular importance since it provides an upper limit for the reliability levels which can be objectively measured in a given amount of observation time purely from observation of failure behaviour. Specialising the equations of §3 to this case is simply a matter of substituting $\langle r_i\rangle = \langle 0\rangle$. If we similarly specialise the form of our *predictions* by considering the Bayesian predictive probability of a *further* period of failure-free operation, we find that these predictions can be expressed in rather a simple form as the expectations of products of higher non-central moments of a particular conditional distribution. So, conclusions about the reliability levels measurable using this model turn out to depend crucially on our decision about what may be considered realistic model assumptions for these moments. Thinking in terms of the probability $Q_i = 1 - P_i$ of successful completion of an individual demand, and assuming that we do believe that our product

family is highly reliable, then the conditional distribution of Q_i given θ will be concentrated very close to 1 (for all except, perhaps, some values of the product-family parameter θ which we consider to be highly unlikely, i.e. that are assigned small probability (density) values $\mathrm{Prior}_\theta(\theta)$ by our prior for θ). Defining μ'_m to be the m^{th} non-central moment of this conditional distribution of Q_i given θ makes μ'_m a deterministic function of θ

$$\mu'_m = \int_0^1 (1-p)^m f_p(p|\theta)\, dp. \tag{15}$$

We now take the expectation of Q_k^n with respect to each of the three updated distributions (12–14) for P_k. This yields three expressions representing the Bayesian predictive probability that the next n demands on $\mathcal{A}_k$ will be failure-free given previous observation of failure-free execution of respectively: $\mathcal{A}_k$ only; $\langle \mathcal{A}_i \rangle_{i=1}^{k-1}$; or, lastly, all of $\langle \mathcal{A}_i \rangle_{i=1}^{k}$:

$$\mathbf{E}(Q_k^n \,|\, R_k = 0) = \frac{\mathbf{E}\big(\mu'_{n_k+n}\big)}{\mathbf{E}\big(\mu'_{n_k}\big)}\,, \tag{16}$$

$$\mathbf{E}\Big(Q_k^n \,\Big|\, \langle R_i \rangle_{i=1}^{k-1} = \langle 0 \rangle\Big) = \frac{\mathbf{E}\left(\mu'_n \prod_{i=1}^{k-1} \mu'_{n_i}\right)}{\mathbf{E}\left(\prod_{i=1}^{k-1} \mu'_{n_i}\right)}\,, \tag{17}$$

$$\mathbf{E}\Big(Q_k^n \,\Big|\, \langle R_i \rangle_{i=1}^{k} = \langle 0 \rangle\Big) = \frac{\mathbf{E}\left(\mu'_{n_k+n} \prod_{i=1}^{k-1} \mu'_{n_i}\right)}{\mathbf{E}\left(\prod_{i=1}^{k} \mu'_{n_i}\right)}\,. \tag{18}$$

These predictive probabilities of n consecutive successful demands on $\mathcal{A}_k$ should be compared with the unconditional

$$\mathbf{E}(Q_k^n) = \mathbf{E}(\mu'_n) \tag{19}$$

which is the probability that the next n demands on $\mathcal{A}_k$ will be failure-free given *no* conditioning observation of either $\mathcal{A}_k$ or any other products—i.e. based solely upon the prior belief.

5 An Example of a Particular Choice of Prior Distributions for P given Θ, and for Θ

We shall retain throughout what follows our original assumptions that each product $\mathcal{A}_i$ fails as a Bernoulli trials process with unknown parameter P_i, and that the $\langle P_i \rangle$ sequence is i.i.d. conditionally given an unknown product-sequence-characterising parameter θ. To generate particular cases of our model

we are then left with the tasks of choosing the distribution family $\{f_p(\cdot|\theta)\,;\,\theta \in \mathcal{S}\}$ and the single prior distribution Prior_θ over this family.

The beta-family of distributions

$$f_p(p|\theta) = \frac{p^{a-1}(1-p)^{b-1}}{\beta(a,b)}, \qquad \theta = \langle a, b\rangle\,, \quad a, b > 0$$

is conjugate [DeGroot 1970] to both the binomial and the negative binomial (including geometric) distributions, and is thus in some sense a 'natural' choice. If we use this as our f_p distribution family, we obtain a mixed process for the failures of a single product for which the probability of r failures in n demands is given from equation (4) to be

$$R\,|\,n, a, b \sim \frac{\binom{n}{r}\beta(r+a, n-r+b)}{\beta(a,b)}\,,$$

obtained by integrating over p the joint distribution of equation (2) which would be

$$(R, P)\,|\,n, a, b \sim \frac{\binom{n}{r}p^{r+a-1}\,(1-p)^{n-r+b-1}}{\beta(a,b)}$$

in this case.

The likelihood (8) resulting from observation of k products in operation is

$$\langle R_i\rangle_{i=1}^k \,\big|\, (\langle n_i\rangle_{i=1}^k\,, a, b) \sim \prod_{i=1}^k \binom{n_i}{r_i} \frac{\beta(a+r_i, b+n_i-r_i)}{\beta(a,b)}$$

with

$$L_k(a,b) = \prod_{i=1}^k \frac{\beta(a+r_i, b+n_i-r_i)}{\beta(a,b)}$$

as defined in equation (9).

Having decided to investigate the beta f_p, the choice of Prior_θ over $\mathcal{S}$, the positive quadrant[4], remains problematic. In real life there would be an 'expert' from whom we would wish to elicit the distribution that truly reflects his a priori belief. This is not an easy task in such a complex model, and the expert may find it difficult to represent his beliefs in a distribution for $\langle a, b\rangle$. A way out of this difficulty is to assume that the expert is 'ignorant', and use that prior distribution which represents ignorance. Even this is a non-trivial task. As an example we consider the simple case of distributions uniform on some finite rectangle with sides parallel to the a and b axes,

$$\text{Prior}_\theta(a,b) = \begin{cases} \frac{1}{(a_2-a_1)(b_2-b_1)}, & \text{if } a_1 < a < a_2,\ b_1 < b < b_2 \\ 0, & \text{elsewhere.} \end{cases}$$

[4] possibly extended to include points representing $a, b \to \infty$ with a/b constant, and $a, b \to 0$ with a/b constant, to include all the limiting cases of the beta family

Firstly we can examine characteristics of the prior distribution (11) for P_k implied by these model assumptions,

$$P_k \sim \int_{a_1}^{a_2}\int_{b_1}^{b_2} \frac{p^{a-1}(1-p)^{b-1}}{\beta(a,b)} \frac{db\,da}{(a_2-a_1)(b_2-b_1)}\,.$$

The first and second non-central moments of $P\,|\,a,b$ are $\frac{a}{a+b}$ and $\frac{a(a+1)}{(a+b)(a+b+1)}$. These may be integrated analytically with respect to $f_p(p|a,b)$ (first expanding in partial fractions with respect to b in the case of the second moment) to give the expressions,

$$\mathbf{E}(P) = \tfrac{1}{2} + \frac{(a_1^2-b_1^2)\log(a_1+b_1) - (a_2^2-b_1^2)\log(a_2+b_1) - (a_1^2-b_2^2)\log(a_1+b_2) + (a_2^2-b_2^2)\log(a_2+b_2)}{2(a_2-a_1)(b_2-b_1)}$$

and

$$\mathbf{E}\big(P^2\big) = \frac{2}{3} + \frac{t(a_1,b_1) - t(a_2,b_1) - t(a_1,b_2) + t(a_2,b_2)}{6(a_2-a_1)(b_2-b_1)}$$

where $t(a,b) = s(a,b) - s(a,b+1)$, where

$$s(a,b) = (2a^2 - 2ab + 2b^2 + 3a - 3b)(a+b)\log(a+b).$$

The prior reliability function is given from equations (15) and (19) by

$$\begin{aligned}\mathbf{P}(X_k > n) &= \mathbf{E}(\mu'_n)\\ &= \int_{a_1}^{a_2}\int_{b_1}^{b_2} \frac{\beta(a,b+n)}{\beta(a,b)} \frac{db\,da}{(a_2-a_1)(b_2-b_1)}\\ &= \int_{a_1}^{a_2}\int_{b_1}^{b_2} \frac{b(b+1)\ldots(b+n-1)}{(a+b)(a+b+1)\ldots(a+b+n-1)} \frac{db\,da}{(a_2-a_1)(b_2-b_1)}\,,\end{aligned}$$

where the *first* failure of $\mathcal{A}_k$ occurs on the X_k^{th} demand.

These expressions can be thought of as different ways of expressing *a priori* belief about the reliability of a product. Now we explore the effects on these beliefs of learning from observation. We examine the realisations under these particular distributional assumptions of both the posterior distributions for P_k given by equations (12–14), and the predictions of X_k, the time to next failure of $\mathcal{A}_k$ using equations (16–18). In the most general case of arbitrary periods of observation of some finite number of previous products, each of the probabilities entailed by these questions takes the form of the ratio of a pair of integrals (over the chosen rectangle in the $\langle a,b\rangle$-plane), where the integrands in the numerator and denominator are each equal to some product of terms of the form

$$\begin{aligned}\mathbf{E}\big(P^r(1-P)^{n-r}\,|\,a,b\big) &= \int_0^1 p^r(1-p)^{n-r}\frac{p^{a-1}(1-p)^{b-1}}{\beta(a,b)}\,dp\\ = \frac{\beta(a+r,b+n-r)}{\beta(a,b)} &= \frac{a(a+1)\ldots(a+r-1)b(b+1)\ldots(b+n-r-1)}{(a+b)(a+b+1)\ldots\ldots\ldots\ldots\ldots(a+b+n-1)}\,.\end{aligned}$$

In practice, since this kind of inference is most likely to be called for in dealing with very high reliability systems, the values n_i of n in these products are

likely to be rather large, and the values of r are likely to be small, and ideally zero. So some very large products will be involved in the above term. We shall report elsewhere on the mathematical difficulties that arise as a result of this. Here we show only some illustrative numerical results based upon the observation of three previous products, each of which has been exposed to 10^7 demands without a single failure. In Table 1 we can see how various different assumptions for Prior_θ affect the strength of the inferences concerning a fourth product in the same family which can be drawn from this sort of evidence of high reliability of previous, similar products.

All the results in the Table involve assuming uniform distributions over different regions of the $\langle a, b\rangle$–space. We have excluded values of b smaller than one, since these entail beta distributions with infinite density at 1; but we have allowed values of a smaller than one, since infinite density at the origin seems plausible. The region in the positive quadrant where a and b are both large can also be ruled out, since any point here corresponds to a beta distribution with very small variance—i.e. it implies that different products will have essentially identical probabilities of failure upon demand, which runs counter to the spirit of this whole exercise.

The first nine rows of the Table involve several rectangles of the kind described above. The ninth row shows a small rectangle, effectively approximating to a known point value for $\langle a, b\rangle$. Rows 10 to 12 show thin 'wedges' adjacent to the b-axis. The informal reasoning here is that it may be reasonable to believe a priori that the mean $\mathbf{E}(P\,|\,a, b)$ of the distribution of probability of failure on demand does not exceed a certain value $0 \leq \mathbf{E}(P\,|\,a, b) \leq M < 1$, say, and this is equivalent to the restriction to $\frac{a}{b} \leq \frac{M}{1-M}$. We used $M = 10^{-3}$, 10^{-5}, and 10^{-7}. Once again, all points in the wedge are given equal weight.

In the Table we show how 'the reliability' of a product changes as a result of the different types of evidence that could be available. For brevity here we have chosen to present the mean of the distribution of P_4, and the reliability function evaluated at 10^7 demands (i.e. the probability of surviving this number of demands), in each of the four cases: given no data; given only evidence of failure-free operation of this product; given only evidence of failure-free working of earlier products; and given both these latter items of evidence.

The most interesting and important results concern the different predictions of future operational behaviour, expressed as the probability $R(10^7)$ of surviving 10^7 further demands without failure: the information from the perfect working of previous products makes only a modest contribution to our confidence in the current product when compared with actual evidence of failure-free working on that product itself (compare columns 8 and 10). Thus when we only have evidence from the previous products, although this is of extensive perfect working for each, it only allows us to claim, in the case of the rectangular priors, about 0.75 probability of similarly extensive perfect working (i.e. surviving 10^7 demands) for the new product[5].

[5]We conjecture that some limiting result may be indicated here : perhaps the probability that product $\mathcal{A}_k$ will survive its first X demands, given that $k-1$ previous products have done so, tends to $(k-1)/k$ as $X \to \infty$.

The evidence from previous perfect working of the *same* product, however, is more informative. It allows us to be much more confident that the product will work perfectly in the future: the probability of it surviving 10^7 demands, given that it has already survived 10^7 demands, exceeds 0.9 in all cases.

On the other hand, the small increase in confidence that comes from experience of other products may be useful in the case of safety-critical systems, especially as it is likely to come with little or no cost to developers of the new product. Thus, in the first row of the Table, the *a priori* belief of the 10^7 demand survival is .062, this increases to .96 after we have actually seen the product survive 10^7 demands, and to .99 when we are told, in addition, that three other products have also survived 10^7 demands. Putting it another way, this evidence of previous product survival has reduced the chance of a failure in the next 10^7 demands by a factor of 4 (from .04 to .01) compared with the result based only on the evidence from operational experience of this product.

We have shown the columns for the means of the various distributions for P_4 mainly as a warning that these can be misleading if used to represent 'the reliability' of a product. Thus the mean probability of failure on demand can be quite large (0.24 in the first line prior distribution), but still the chance of surviving 10^7 demands may be non-negligible (0.063 in this case). The informal reason is that the distribution is such that the mean is not a good summary statistic, and in particular cannot be used in a geometric distribution to approximate to the more complex model that applies here.

In fact, decreasing values of $\mathbf{E}(P_4)$ do not necessarily imply increasing chance of surviving 10^7 demands, as might naively be expected: see, for example, columns 7 and 8 of rows 1 to 4. Imagine that we have two experts, let us call them James and Peter, represented by two different prior distributions (rows of the Table), who observe the system to survive for 10^7 demands. They are then asked to tell us how reliable the system is. If the question is posed as 'what is the mean of P_4?', then James is more optimistic than Peter; if, however, the question is posed as 'what is the chance of surviving a further 10^7 demands', Peter is more optimistic than James. Such (only apparent) paradoxes underline the importance of using the right formulation for our purposes when we ask questions about the reliability of a system.

6 Conclusions and future work

A major motivation for research of this kind is to make the process of assessing safety-critical systems more open to analysis. Currently, particularly in those cases where complex software is involved, such assessments have a high degree of informality and rely a great deal upon expert judgement. Whilst this process is usually carried out responsibly, and with great rigour, it is difficult for an outsider to analyse how the final judgement has been reached, and much has to be taken on trust. Since there is some evidence of experts being unduly optimistic about their judgemental abilities [Henrion, Fischhoff 1986], simply checking their honesty is insufficient. What is needed is a more formal means

of argumentation, where the assumptions and reasoning processes are visible and can be questioned. This new model treats a small part of this problem by providing a representation, and means of composition, of two important types of evidence that are commonly used to make claims for the reliability of a product: evidence from testing of the product itself and evidence from previous experience of 'similar' products.

Whilst we make no great claims for the realism of the example we have used, it does indicate the way in which a formal model of this kind could be used to question whether an optimistic conclusion drawn from past experience might be ill-founded. Essentially, if you were to claim that great trust could be placed in a particular system because of past experience of other systems, you would have to justify this by trying to claim that your prior distribution is reasonable within the model. It is clear that some of the examples of prior distributions we have used could be said to be 'unreasonable' in the sense that they represent beliefs about the reliability, prior to seeing any evidence, that are very strong.

The particular numerical examples used here are meant only to be illustrative. Clearly further work is needed to identify classes of 'plausible' prior distributions, even for the case in which the expert professes 'complete prior ignorance'. For example, rather than addressing the raw $\langle a, b \rangle$ parameters, it may be easier for the subject to think in terms of a reparameterisation - the mean and coefficient of variation are possibilities. Another area of future work concerns the impact of different kinds of evidence upon the conclusions. For example, the case here of complete perfection of operation of the previous products is the best news that it is possible to have, and it would be interesting to investigate the case where there have been failures in the earlier products.

The possibility that conclusions about the reliability of a system can be highly dependent upon the precise way in which they are formulated is somewhat surprising and needs further investigation. However, the results here support those obtained in a different context, concerning stopping rules for software testing [Littlewood, Wright 1995].

Finally, all this modelling depends upon the reasonableness of notions of 'similarity' between different products. In this we are merely making more formal the extremely informal claims that experts make when they argue that the behaviour of one product can be used as a means of inferring the likely behaviour of another. Justification of such assumptions of similarity in particular cases is, of course, outside the direct scope of our studies—presumably it will come, in the case of software, from knowledge of the application domain (the problems being solved were similar), the development process (the methods used were similar), the design teams (they were the same or of comparable competence), etc. However, we believe that our model can be used to provide a curb on the enthusiasm of experts: specifically, the use of 'similarity' arguments to make stronger claims than would be warranted via the model should be treated with suspicion.

Region of Uniform Prior				Given no Data		Given no failure of this product		Given no failure of previous 3 products		Given failure neither of this nor of previous 3 products	
a_1	a_2	b_1	b_2	$\mathbf{E}(P_4)$	$R(10^7)$	$\mathbf{E}(P_4)$	$R(10^7)$	$\mathbf{E}(P_4)$	$R(10^7)$	$\mathbf{E}(P_4)$	$R(10^7)$
0	1	1	2	.2384	.6229E-1	.3966E-1	.9585	.1388E-1	.7498	.1047E-1	.9893
0	1	1	10	.1037	.6828E-1	.1577E-1	.9547	.5398E-2	.7499	.4062E-2	.9883
0	1	1	100	.2077E-1	.8048E-1	.3020E-2	.9469	.1019E-2	.7500	.7655E-3	.9862
0	1	1	1000	.3207E-2	.9877E-1	.4636E-3	.9355	.1556E-3	.7500	.1168E-3	.9831
0	2	1	2	.3692	.3114E-1	.3966E-1	.9585	.1388E-1	.7498	.1047E-1	.9893
0	2	1	10	.1781	.3414E-1	.1578E-1	.9547	.5398E-2	.7499	.4062E-2	.9883
0	2	1	100	.3833E-1	.4024E-1	.3020E-2	.9469	.1019E-2	.7500	.7655E-3	.9862
0	2	1	1000	.6091E-2	.4939E-1	.4637E-3	.9355	.1556E-3	.7500	.1168E-3	.9831
.01	.0101	10	10.1	.9990E-3	.8700	.9990E-3	.9931	.9990E-3	.8700	.9990E-3	.9931
0	$b/999$	1	1000	.5002E-3	.1824	.2056E-3	.9401	.9494E-4	.7545	.7593E-4	.9832
0	$b/99999$	1	1000	.5000E-5	.9689	.4947E-5	.9977	.4843E-5	.9703	.4791E-5	.9978
0	$b/9999999$	1	1000	.5000E-7	.99968	.4999E-7	.999977	.4998E-7	.99968	.4998E-7	.999977

.XXXXE-n means $0.XXXX \times 10^{-n}$

Table 1: Effect on Reliability Predictions of Observation of Non-Failure of Previous Products

Acknowledgement

This work was supported by the ESPRIT PDCS2 project 6362, the DTI/ EPSERC Safety Critical Systems Research Programme's DATUM Project, and the CEC Environment Programme's SHIP Project. It has benefited considerably from numerous critical comments and suggested improvements by colleagues working on these projects and colleagues at the Centre for Software Reliability.

References

[Butler, Finelli 1993] R. W. Butler and G. B. Finelli. The infeasibility of quantifying the reliability of life-critical real-time software. *IEEE Transactions on Software Engineering*, 19(1):3–12, 1993.

[DeGroot 1970] M. H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

[Henrion, Fischhoff 1986] M. Henrion and B. Fischhoff. Assessing Uncertainty in Physical Constants. *American Journal of Physics*, 54(9):791–8, 1986.

[Laprie 1992] J. C. Laprie. For a Product-in-a-Process Approach to Software Reliability Evaluation. In *Proc. 3rd International Symposium on Software Reliability Engineering (ISSRE92)*, pages 134–9, Research-Triangle Park, USA, 1992. Invited Paper.

[Littlewood, Strigini 1993] B. Littlewood and L. Strigini. Validation of Ultra-High Dependability for Software-Based Systems. *Comm. Assoc. Computing Machinery*, 36(11):69–80, November 1993.

[Littlewood, Wright 1995] B. Littlewood and D. Wright. Stopping Rules for the Operational Testing of Safety-Critical Software. In 25^{th} *Fault Tolerant Computing Symposium (Pasadena, California), Digest of Papers*, pages 444–51. IEEE, IEEE Computer Society Press, 1995.

[Rouquet, Traverse 1986] J. C. Rouquet and Z. Z. Traverse. Safe and Reliable Computing on board the Airbus and ATR aircraft. In W. J. Quirk, Editor, *Proc. Fifth IFAC Worshop on Safety of Computer Control Systems*, pages 93–97, Oxford, 1986. Pergamon Press.