Reliability Prediction of a Software Product Using Testing Data from other Products or Execution Environments

DeVa Project Technical Report 10

Bev Littlewood and David Wright

City University
Northampton Square, London EC1V 0HB

2nd December 1996

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. The model we propose could equally be applied to increase our confidence that a product will operate reliably in a novel environment, using evidence of its past behaviour in previous environments. We present some illustrative numerical results that seem to suggest that such experience from previous products or environments, even where very high operational dependability has been achieved, can only modestly improve our confidence in the reliability of a new product or of an existing product when transferred to a new environment.

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, [2, 7, 9]. 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, subject to reasonably plausible assumptions, 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 [11, 10], 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. This seems to be the way in which some safety-critical software-based systems are currently assessed, although it must be said that the process of combining evidence here is somewhat informal and does not generally provide a *quantitative* assessment [5]. The different kinds of potential evidence 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 within a particular environment by also taking into account the data on success/failure sequences either of other products or of the current product executing under operational conditions which differ from the present conditions. Thus these other failure data sets may be records of the success (or not) in building and operating 'similar' products in the past. Alternatively, they may originate as records of execution of the current product in different environments. An important special case, of course, is the one where there is unreserved good news from these previous data sets—i.e. where there have been no failures in any of the data sets 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 within an operating environment. 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.

The model contained here may be applicable either to the data sets arising from a number of different software products, or from a single software product executing in a number of different operational environments. The 'indifference' assumption discussed below is all that is required in either case. Thus we can think of our 'experimental unit' as a particular product operating in a particular environment. From each such $\langle \text{product}, \text{environment} \rangle$ unit, we observe operational data. Since we have chosen to work in terms of discrete time—a sequence of discrete demands on the software, each resulting in a sequence of successes and failures—the operational data arising from each $\langle \text{product}, \text{environment} \rangle$ unit is a binary sequence. We are interested in using data from a family of such units to improve our ability to forecast a particular one of them, i.e. a

particular one of the binary success-failure sequences. Here, we make what we see as the simplest assumption which allows this kind of learning from one sequence to another: An assumption of prior indifference between the members of our family of operational success-failure sequences. For example, we assume that we are without prior[1] beliefs of any kind which would cause us to identify some particular pair $\mathcal{A}$ and $\mathcal{B}$ of testing data sequences about which we could say 'I expect that sequence $\mathcal{A}$ will show greater reliability than sequence $\mathcal{B}$'. Our model says that such beliefs about comparisons of reliabilities between different sequences will emerge only after we begin to examine the numerical values of the failure-counts which those sequences contain. With this understanding of our meaning, we refer in what follows to a set of success-failure sequences about which we feel this indifference as a family of 'similar' sequences (emanating from a family of 'similar' $\langle$product, environment$\rangle$ pairs). So in our usage here the 'similarity' of the success-failure sequences within a family is nothing more than a prior statistical indifference between these sequences. Of course, this idea might elsewhere be extended by means of an ordering of the distinct sequences and some kind of process model for, say, increasing reliability expectations from one $\langle$product, environment$\rangle$ to the next. But our indifference assumption is simpler, whilst allowing us to explore mathematically the learning which might take place from one sequence to the next and, we believe, being a plausible model in some circumstances. In particular, even this simple model well illustrates the importance of prior belief—about the statistical relationship between these failure-success sequences—for any conclusions we might wish to draw from data on other products or operating environments.

In the next section a doubly stochastic Bayesian model of the failures (if any) of a family of 'similar' software success-failure sequences is constructed. The intention is to augment the relatively meagre evidence that can realistically be gained from testing of a particular product in a single environment. We can now take account also of the success (or not) in conducting similar operational trials in the past. The analytical results which follow in Section 3 lead to an examination of an important special case in Section 4. Section 4 explores the conclusions which can legitimately be drawn from observation of a number of sequences all of which contain no failure up till the present time. We examine this no-past-failure case in some detail, and, after a brief enumeration in Section 5 of some practical questions whose answers our model might be used to explore, we proceed in Section 6, to obtain some algebraic and numerical results for a few example cases of our general model. In discussing these special cases which arise from introducing specific parametric distributional assumptions, we concentrate mainly on the no-failures case introduced in Section 4. Some considerations about the difficulties of choosing a measure of reliability are mentioned in Section 7. Our main conclusions are summarised in Section 8. The appendices contain some of the mathematical details required for the main part of the report, including, in Appendix C our procedure for calculating very high order non-central moments of the Beta distribution which we used for the numerical work of Section 6.2.

2 Modelling Approach

We wish to use evidence we have obtained from building and operating previous products, or from previous operational use (in different environments) of our current product, in order to try to improve the accuracy of the predictions that we can make about the reliability either of an entirely novel product or a previously used product which now operates in a novel environment. To do this we must take account of two kinds of uncertainty. In the first place, there will be uncertainty concerning the actual reliabilities that were achieved by these earlier $\langle$product, environment$\rangle$ applications. Even in those cases where there is extensive operating experience, we shall never know the true reliability of a given product in a given environment and will have to use an estimate based upon the finite amount of operational data collected during its use within the environment concerned. In those situations where we are dealing with products that are likely to be very reliable in their intended environments, 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

[1] prior to observing the success-failure data itself

concern the statistical 'similarity' to one another of the success-failure sequences that have been observed in the past and the 'similarity' of the one under study to these past sequences. Clearly it will be misleading (and give optimistic results) if we simply assume these earlier sequences, and the present one, are 'exactly similar' in the sense that they all arise from exactly the same true reliability [6].

In what follows, we shall assume that the true per-demand probabilities of failure of the different sequences, 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 sequences 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, and the operational environments in which they are situated 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 success-failure sequences 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 of two $\langle \text{product}, \text{environment} \rangle$ pairs $\mathcal{A}$ and $\mathcal{B}$ would show the best reliability, you would have no preference. This is represented by their probabilities of failure per single demand being identically distributed random variables: any probability statements you would make about the reliabilities of demand sequences $\mathcal{A}$ and $\mathcal{B}$ would 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 sequences will show identical reliabilities - indeed you will know that the actual reliabilities of the sequences will differ. The two sources of uncertainty here are both important. However, it is the nature of the uncertainty concerning 'how similar' the sequences actually are that will be most difficult to estimate in practice, since this requires us to see as many different sequences as possible. That is, we would require operational data on a large number of $\langle \text{product}, \text{environment} \rangle$ pairs between any two of which, prior to inspection of actual failure data, we felt indifferent. But in practice, it is far more likely that we have large quantities of testing information about a few $\langle \text{product}, \text{environment} \rangle$ combinations than it is that we have information on many such testing sequences.

Consider first the failure process of a *single* software success-failure sequence $\mathcal{A}$. Assume a Bernoulli trials process model of the failures of this $\langle \text{product}, \text{environment} \rangle$ in a sequence of 'demands' with neither debugging, maintenance, nor significant variation in the 'stressfulness' of the software's operational environment. An example might be the installed software protection system of a nuclear reactor, where demands could be assumed to be sufficiently separated in time as to be treated as independent. For flexibility of expression we will use $\mathcal{A}$ to refer to both the $\langle \text{product}, \text{environment} \rangle$ pair and the 'sequence' of successes and failures on successive demands on this software in this environment. Then strictly, 'sequence' means the exact *probability law*[2] *governing the sequence*, rather than the realised boolean values of the sequence. With this understanding, we can refer to $\mathcal{A}$ sometimes as a 'pair', at other times as a sequence, and even as a single number p which we have an interest in estimating as accurately as possible. Thus, in the first n trials of sequence $\mathcal{A}$, 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 binomial:

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

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, possibly a vector. Here we can think of the shape of this distribution $f_p(p|\theta)$ for P as a representation of the general reliabilities of sequences in a particular *family of*

[2] by assumption in our model a Bernoulli trials process completely specified by a single numerical parameter p

$\langle$*product, environment*$\rangle$ *pairs*, perhaps representing the different failure histories of a single product executing in multiple environments. Alternatively this family might consist of the failure histories of a number of 'similar' products produced by a single development team, using a common development method, and for similar applications. For example, a family of data sequences known to have highly variable reliability levels would correspond to a distribution $f_p(p|\theta)$ with a large variance, whereas, for another family of sequences, an expected high 'average' reliability figure would correspond to a small mean for $f_p(p|\theta)$. If we fully understood the true variation in reliabilities of the sequences in each of these two success-failure sequence 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 $\langle$product, environment$\rangle$ pairs. That is, given sufficient data on the reliability variation amongst the sequences 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 parameter characterising a family of $\langle$product, environment$\rangle$ pairs. For a $\langle$product, environment$\rangle$ chosen at random from those of a particular family (i.e. particular θ) and observed for the first n demands, it follows that (R, P) has joint distribution[3]

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

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 $f_p(\cdot|\theta)$ (a form which will repeatedly be found in results later),

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

Note from (3) that mixing over p using this fixed θ, not surprisingly, has the effect that the distribution for the number of failures which will be seen during a given sequence of demands is now *more dispersed* than a corresponding binomial distribution. We can quantify this effect precisely by verifying that from the distribution (3) we have mean

$$\mathbf{E}(R\,|\,n, \theta) = n\mathbf{E}(P\,|\,\theta)$$

where

$$\mathbf{E}(P\,|\,\theta) = \int_0^1 p f_p(p|\theta)\,dp\,,$$

and

$$\mathbf{Var}(R|n, \theta) = n\mathbf{E}(P\,|\,\theta)\,(1 - \mathbf{E}(P\,|\,\theta)) + n(n-1)\mathbf{Var}(P\,|\,\theta)\ .$$

In this sum, the left-hand term is the variance of a binomial distribution with the same maximum n and mean $n\mathbf{E}(P\,|\,\theta)$. As one might expect, the right-hand 'excess' term depends on the variance

$$\mathbf{Var}(P\,|\,\theta) = \int_0^1 (p - \mathbf{E}(P\,|\,\theta))^2 f_p(p|\theta)\,dp$$

of the mixing distribution $f_p(\,\cdot\,|\theta)$.

[3] Notice that we keep to the usual notational convention of upper case for a random variable and lower case for a numerical value obtained as a particular realisation.

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 [3, pp213-4,217], where $f_p(\cdot|\theta)$ is the mixing distribution. Note that although exchangeability[4] 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 from the same sequence now has an updated distribution obtainable from (5) as

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

given n, r.

The distributions which we have considered up till this point are parameterised by θ. Under our chosen model, (6) is not a predictive distribution of future failures given past failure behaviour since it depends on the unknown value of the parameter θ. This deliberate θ-dependence is intended to take account of the practical fact that we are unable with any confidence to accurately state the distribution $f_p\,(\,\cdot\,|\theta)$ of failure probabilities of the $\langle$product, environment$\rangle$ pair within our family. This inability is captured in the model as our uncertainty about the $\langle$product, environment$\rangle$-family parameter θ. This parameter uncertainty has yet to be expressed and incorporated into the picture. We now adopt a Bayesian approach to handling this dimension of the problem 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 $\langle$product, environment$\rangle$, this extension actually adds very little, if anything, useful 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.) However, the idea of a prior distribution for θ becomes a useful concept if we wish to address the problem of *learning* about a *distribution* of software reliabilities by observing *multiple* sequences of software failure behaviour from a single family $\langle\mathcal{A}_i\rangle$, say, of $\langle$product, environment$\rangle$ pairs. We can then represent a conservative[5] version of a *process* concept for their reliabilities, from one $\langle$product, environment$\rangle$ to the next, by modelling these sequences as individual failure processes of the Bernoulli-trials kind discussed above but 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 $\langle$product, environment$\rangle$ pairs. We are then able to learn from observation of the early data sequences about the likely behaviour of another sequence through the medium of our improving knowledge of their common parameter θ.

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 testing sequence, θ now represents a common unknown

[4] The property that any permutation of a portion of the boolean (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.

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

characteristic of the whole family of such sequences. To obtain the value of θ would be to capture the reliability-relevant characteristic which these software pairs $\langle$product, environment$\rangle$ all have in common. For this *multi*-sequence model, there is now a real purpose behind including separate distributional assumptions for firstly θ, and secondly p_i given θ. Below, we do not in fact assume that θ can ever be known[6]. 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 sequences 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 family of sequences is a single unknown realisation. Figure 1 depicts these conditional dependence relationships diagramatically. This popular DAG (directed acyclic graph) representation of conditional independence assumptions is equivalent to the assertion that the joint distribution of all the nodes is equal to the product of the conditional distributions of each node conditioned on the values of its parents. Actually, we have tended to condition on the values $\langle n_i\rangle_{i=1}^{k}$ throughout our probabilistic analysis so that the n_i-nodes can be thought of as degenerate, constant random variables. Note that we have used a notation for our mixtures and marginal distributions

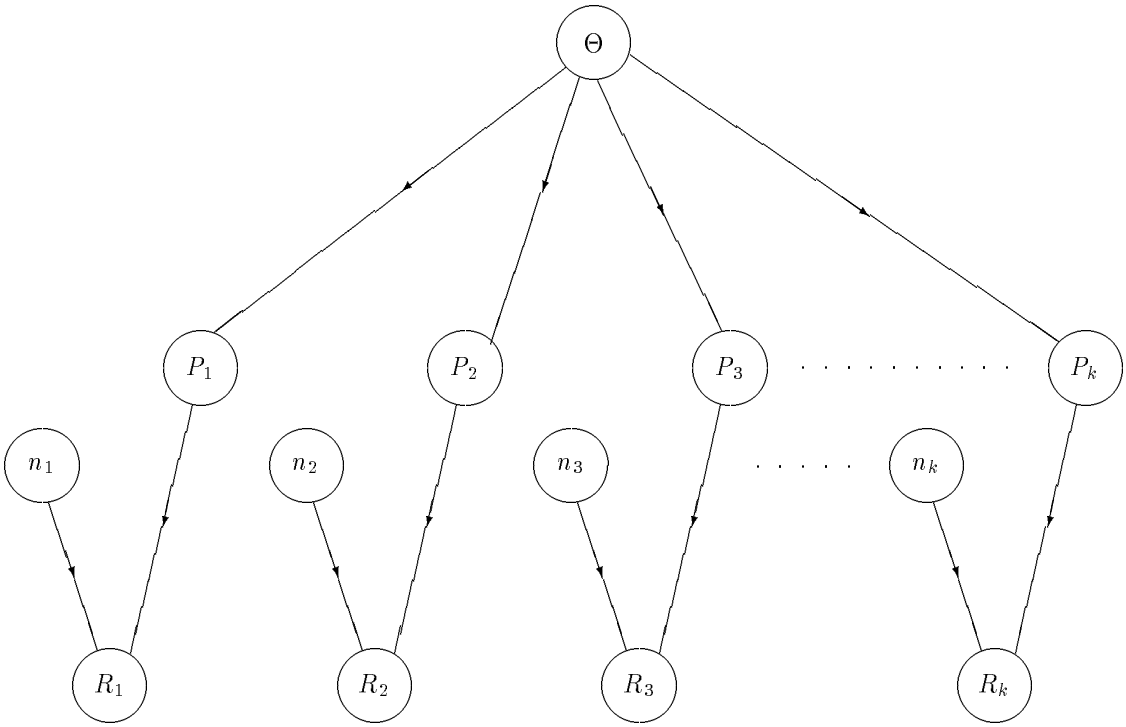


Figure 1: Diagram of the dependencies of the model

which assumes that both the distributions f_p, for P given θ, and the prior distribution, Prior_θ for θ, are continuous. The cases where either or both of these distributions are discrete are also of interest and correspond to the replacement of integrals by sums, or, alternatively, to the use of the Dirac delta function in specifying definitions for our densities f_p and Prior_θ.

Before proceeding to consider in §6 specific distributional assumptions appropriate for the

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

i.i.d. P_i given Θ, and for the $\langle$product, environment$\rangle$ family parameter Θ itself, we obtain, in the following sections, a few consequences of these model assumptions in the general case. Observe firstly that, conditionally given θ and $\langle n_i \rangle_{i=1}^{k}$, our independence assumption for the $\langle P_i \rangle$ tells 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 are in possession of data in the form of observed failure behaviour of these k software products executing in their k environments (i.e., r_i failures out of n_i trials for each sequence $\mathcal{A}_i$) 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 constant[7] combinatorial terms together with moments of the parametric distribution $f_p(\cdot|\theta)$

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

We find, not surprisingly, in the following sections that, using this Bayesian model, our reliability predictions turn out to depend heavily on our *prior beliefs*, and not only on the empirical reliability data $\langle r_i \rangle_{i=1}^{k}$ which is later observed. We have expressed the shape of these beliefs formally by our selection of the distributions $\langle \{f_p(\,\cdot\,|\theta); \theta \in \mathcal{S}\}, \mathrm{Prior}_\theta \rangle$ comprising our model for the failure probabilities $\langle P_i \rangle$ of our family $\langle \mathcal{A}_i \rangle$ of sequences of Bernoulli trials. There are several ways of understanding the entity $\langle \{f_p(\,\cdot\,|\theta); \theta \in \mathcal{S}\}, \mathrm{Prior}_\theta \rangle$ less formally, which may help with the selection of appropriate distributions in the case of a particular family $\langle \mathcal{A}_i \rangle$. To begin with, $\langle \{f_p(\,\cdot\,|\theta); \theta \in \mathcal{S}\}, \mathrm{Prior}_\theta \rangle$ should contain at least

1. Our best guess, prior to observation, of the *family average reliability level* of the $\langle \mathcal{A}_i \rangle$,

i.e. the average reliability towards which our beliefs would hypothetically converge if we could acquire arbitrarily large amounts of data from each of arbitrarily many distinct $\langle$product, environment$\rangle$ pairs which were representative of this family. But $\langle \{f_p(\,\cdot\,|\theta); \theta \in \mathcal{S}\}, \mathrm{Prior}_\theta \rangle$ is much more than a complicated way of expressing a guess at the family average reliability. We emphasise that it contains at least two other dimensions of expressed prior belief, each of which should be verified against intuition, and against any available objective prior knowledge, if this model is to be applied. $\langle \{f_p(\,\cdot\,|\theta); \theta \in \mathcal{S}\}, \mathrm{Prior}_\theta \rangle$ also contains

2. Our prior beliefs about the *shape of the distribution of true reliabilities* (as distributed around the average value assessed in 1.) of this $\langle \mathcal{A}_i \rangle$ family. How *consistent* will the reliabilities governing success and failure in the testing sequences $\mathcal{A}_i$ of our family eventually be found to be?

Lastly, but of no lesser significance in terms of the reliability predictions emanating from our approach, we recall that the entity $\langle \{f_p(\,\cdot\,|\theta); \theta \in \mathcal{S}\}, \mathrm{Prior}_\theta \rangle$ incorporates a Bayesian subjective parameter distribution Prior_θ. Through this, the construction $\langle \{f_p(\,\cdot\,|\theta); \theta \in \mathcal{S}\}, \mathrm{Prior}_\theta \rangle$ pays due regard to our stated measure of

3. Our *confidence* (or lack of it) in own our *ability to produce accurate a priori guesses* at 1. and 2. How confident are we that both of these initial assessments are close to the truths that would ultimately be discovered given unlimited data from an unlimited number of representative boolean valued sequences $\mathcal{A}_i$?

This third component of prior belief is a classic Bayesian subjective prior distribution describing our uncertainty about a model feature which in this case is effectively an entire continuous probability distribution on the unit interval, and whose unknown true value characterises our whole family of $\langle$product, environment$\rangle$ pairs. This is perhaps also the component whose effects on the subsequent analysis are the most easy to overlook—or at least whose effects in our analysis can be

[7] i.e. not depending on θ

the most difficult to follow intuitively. To simplify for the sake of illustration, suppose we make θ a one-dimensional real quantity (so that $\mathcal{S} \subseteq \mathbb{R}$), and suppose that we happen to have used a parameterisation of our family of $f_p(\cdot|\theta)$ distributions that orders these distributions according to their means. Then, holding this parameterisation $\{f_p(\cdot|\theta); \theta \in \mathcal{S}\}$ fixed, the act of choosing a relatively more dispersed distribution Prior_θ will correspond to a statement of relatively lower confidence in our ability to accurately guess the value of the family average of reliabilities (item 1. above). Similarly, if we adopt, say, the coefficient of variation of the unknown true distribution from which the P_i are drawn as a numerical representation of an important attribute of the shape (item 2. above) of this distribution of failure probabilities P_i around this mean value, and if we assume, again for simplicity, that our chosen θ-parameterisation now orders the distributions $f_p(\cdot|\theta)$ instead according to their coefficients of variation rather than their means, then the choice of a relatively more dispersed Prior_θ will represent our relatively lower confidence in our ability to assess, a priori, the true amount of consistency amongst the different reliabilities of the members of our family $\langle \mathcal{A}_i \rangle$. So in the case of such a parameterisation we could ask ourselves whether we already possess a thorough understanding of reliability variability within this kind of $\langle \text{product}, \text{environment} \rangle$ family. If so, then a highly concentrated Prior_θ distribution would be an appropriate choice. If instead we considered within-family reliability variation between the $\mathcal{A}_i$ to be rather difficult to assess, without spending some time accumulating an operational history of a number of sequences from the family concerned, then we should choose a larger spread for Prior_θ : and by so doing admit a greater variety of distributions $f_p(\cdot|\theta)$ on the unit interval which could each plausibly represent the true nature of the variation of failure probabilities between the sequences of our family.

Of course, we do not have to use the exact 3 items defined above in order to informally decompose the structure $\langle \{f_p(\cdot|\theta); \theta \in \mathcal{S}\}, \text{Prior}_\theta \rangle$ of our prior belief into a number of salient features whose effects will transmit themselves through the mathematical analysis of this model. The important point is that we must be aware of the profound implications—for the reliability predictions obtained in the following sections—that each one of these components of our prior belief model has. To summarise the last paragraph, the model introduced in this paper proposes a formal representation of prior beliefs about a family of $\langle \text{product}, \text{environment} \rangle$ pairs between which we are initially indifferent. This representation $\langle \{f_p(\cdot|\theta); \theta \in \mathcal{S}\}, \text{Prior}_\theta \rangle$ is expressive enough to allow us—in fact it requires us—to state with precision *how much we know* (and often the extent to which we are in fact ignorant) about the average *level* and *the distribution* of the achieved reliabilities of the members of this family. Given an available amount $\langle k, n_1, n_2, \ldots, n_k \rangle$ of testing data, we go on in the following sections to show how this model of prior belief, combines with empirical testing data $\langle r_1, r_2, \ldots, r_k \rangle$ to yield predictions of future reliability of an individual success-failure sequence within the family. An important question running through the analysis of this model is the amount of improvement in our ability to assess high reliabilities that is achieved by incorporating data on other sequences within the family. It is of interest to examine formally the dependence of the answer to this question on the strength of our prior beliefs—particularly our prior beliefs about reliability consistency—concerning the family $\langle A_i \rangle$.

3 Bayesian Updating of Distributions and Moments in the General Case

To implement the Bayesian learning about Θ given observation of $\langle r_i \rangle_{i=1}^k$ we need to calculate the posterior distribution of Θ

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

where c is a function of $\langle r_i, n_i \rangle_{i=1}^{k}$ not involving θ, i.e.

$$\Theta \,\big|\, \langle n_i, r_i \rangle_{i=1}^{k} \sim \frac{\left[\prod_{i=1}^{k} \int_0^1 p^{r_i} \, (1-p)^{n_i - r_i} \, f_p(p|\theta) \, dp\right] \mathrm{Prior}_\theta(\theta)}{\int_{\theta \in \mathcal{S}} \left[\prod_{i=1}^{k} \int_0^1 p^{r_i} \, (1-p)^{n_i - r_i} \, f_p(p|\theta) \, dp\right] \mathrm{Prior}_\theta(\theta) \, d\theta} \tag{9}$$

Equation (9) draws the focus of attention away from failure probabilities P_i of sequences $\mathcal{A}_i$ by the integrations over p. But it is now of great practical interest to know an up-to-date distribution for P given what has been observed (in order to make predictions about a particular new $\langle \mathrm{product}, \mathrm{environment} \rangle$, 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 sequence. At this stage it is instructive to distinguish three different circumstances under which we will have learned, in different ways, about one of the failure probabilities, say P_k. These three different circumstances will each result in an up-to-date Bayesian posterior distribution for this failure probability, which may be compared with the prior marginal distribution of P_k

$$P_k \sim \int_{\theta \in \mathcal{S}} f_p(p_k|\theta) \mathrm{Prior}_\theta(\theta) \, d\theta \,, \tag{10}$$

which represents our initial state of uncertainty concerning the reliability of any given sequence, $\mathcal{A}_k$, prior to any observation either of that or of any other $\langle \mathrm{product}, \mathrm{environment} \rangle$ pair's behaviour. At this point of no observation, (10) is the mixing distribution associated with our mixture-of-Bernoulli-trials model for future failure of $\mathcal{A}_k$. This comparison of (10) with subsequent updated P_k-distributions determines the nature and limits of what we can learn from observed failure behaviour alone, be it of a single sequence or of a number of sequences from a particular $\langle \mathrm{product}, \mathrm{environment} \rangle$ family.

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

$$P_k \,|\, n_k, r_k \sim \frac{p_k^{r_k} (1-p_k)^{n_k - r_k} \int_{\theta \in \mathcal{S}} f_p(p_k|\theta) \mathrm{Prior}_\theta(\theta) \, d\theta}{\int_{\theta \in \mathcal{S}} \int_0^1 p^{r_k} \, (1-p)^{n_k - r_k} \, f_p(p|\theta) \, dp \, \mathrm{Prior}_\theta(\theta) \, d\theta} \tag{11}$$

for P_k given n_k and r_k. Note that we will continually assume, as we have done in the denominator here, that the families of densities chosen are such that changes of the order of integration are legitimate.

Secondly replacing k by $k-1$ in (9) and then substituting this distribution in place of $\mathrm{Prior}_\theta(\theta)$ in (10) (or, alternatively, directly substituting $n_k = r_k = 0$ in (13) below) gives the distribution

$$P_k \,\big|\, \langle n_i, r_i \rangle_{i=1}^{k-1} \sim \frac{\int_{\theta \in \mathcal{S}} f_p(p_k|\theta) \left[\prod_{i=1}^{k-1} \int_0^1 p^{r_i} \, (1-p)^{n_i - r_i} \, f_p(p|\theta) \, dp\right] \mathrm{Prior}_\theta(\theta) \, d\theta}{\int_{\theta \in \mathcal{S}} \left[\prod_{i=1}^{k-1} \int_0^1 p^{r_i} \, (1-p)^{n_i - r_i} \, f_p(p|\theta) \, dp\right] \mathrm{Prior}_\theta(\theta) \, d\theta} \tag{12}$$

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

Thirdly, replacing k by $k-1$ in (9) and then substituting this distribution in place of $\mathrm{Prior}_\theta(\theta)$ in (11) gives the distribution

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

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

We remark here that the effects of observation respectively of past and of the present software $\langle$product, environment$\rangle$ pair's failure behaviour on our beliefs about the present pair's per-demand failure probability P_k appear to obey a simple multiplicative property. Comparing the *numerators* of the four different probability densities of P_k given by equations (10–13) we see that these are in common proportions to each other[8]. The denominators appear to spoil these relationships, but the denominators are only normalising constants, i.e. they do not depend on P_k. We can use this fact to express the property more concisely in terms of the effect of the different observations on the extent to which we favour one value, say p'_k, of P_k over another, say p''_k. If, in this way, we compare the values of the densities at this arbitrary pair of P_k values, we see that

$$\frac{\mathrm{pdf}\left(p'_k \left| \langle n_i, r_i \rangle_{i=1}^{k} \right.\right)}{\mathrm{pdf}\left(p''_k \left| \langle n_i, r_i \rangle_{i=1}^{k} \right.\right)} \cdot \frac{\mathrm{pdf}\,(p'_k)}{\mathrm{pdf}\,(p''_k)} = \frac{\mathrm{pdf}\,(p'_k \,|n_k, r_k)}{\mathrm{pdf}\,(p''_k \,|n_k, r_k)} \cdot \frac{\mathrm{pdf}\left(p'_k \left| \langle n_i, r_i \rangle_{i=1}^{k-1} \right.\right)}{\mathrm{pdf}\left(p''_k \left| \langle n_i, r_i \rangle_{i=1}^{k-1} \right.\right)} \tag{14}$$

provided, of course, that p''_k is not a point of zero density for any of these four densities. Equation (14) perhaps becomes more intuitively meaningful if converted to the form

$$\frac{\mathrm{pdf}\left(p'_k \left| \langle n_i, r_i \rangle_{i=1}^{k} \right.\right)}{\mathrm{pdf}\left(p''_k \left| \langle n_i, r_i \rangle_{i=1}^{k} \right.\right)} \Big/ \frac{\mathrm{pdf}\,(p'_k)}{\mathrm{pdf}\,(p''_k)} = \left\{ \frac{\mathrm{pdf}\,(p'_k \,|n_k, r_k)}{\mathrm{pdf}\,(p''_k \,|n_k, r_k)} \Big/ \frac{\mathrm{pdf}\,(p'_k)}{\mathrm{pdf}\,(p''_k)} \right\} \cdot \left\{ \frac{\mathrm{pdf}\left(p'_k \left| \langle n_i, r_i \rangle_{i=1}^{k-1} \right.\right)}{\mathrm{pdf}\left(p''_k \left| \langle n_i, r_i \rangle_{i=1}^{k-1} \right.\right)} \Big/ \frac{\mathrm{pdf}\,(p'_k)}{\mathrm{pdf}\,(p''_k)} \right\}. \tag{15}$$

In practical terms this says that observation of *both the present and previous sequences* changes the 'odds of $P_k = p'_k$ vs. $P_k = p''_k$' by a factor which is the product of the corresponding changes in odds resulting from observing, respectively, *only the present* sequence, or *only previous* sequences[9]. The same property is alternatively captured by the formula

$$P_k \left| \langle n_i, r_i \rangle_{i=1}^{k} \sim \frac{\mathrm{pdf}\,(p_k|n_k, r_k) \cdot \mathrm{pdf}\left(p_k \left| \langle n_i, r_i \rangle_{i=1}^{k-1} \right.\right) / \mathrm{pdf}\,(p_k)}{\int_0^1 \mathrm{pdf}\,(p_k|n_k, r_k) \cdot \mathrm{pdf}\left(p_k \left| \langle n_i, r_i \rangle_{i=1}^{k-1} \right.\right) / \mathrm{pdf}\,(p_k) \; dp_k} \right. \tag{16}$$

(defining this density to be zero wherever $\mathrm{pdf}\,(p_k)$ is zero).

On closer examination the model property captured by (14) and (15) is found to be merely an instance of a quite general result in Bayesian statistical modelling which applies wherever the construction of a probability model makes two observables, Y_1 and Y_2, *conditionally* independent *given* the value of some model parameter ξ. Then if we ask the question: 'How do the Bayesian updated distributions of ξ for the three possible cases relative to observation or non-observation of Y_1 and Y_2 compare with the prior distribution of ξ?', we obtain an answer of the above form. Our model is clearly of this kind for $\xi = p_k$, $Y_1 = \langle R_i \rangle_{i=1}^{k-1}$, and $Y_2 = R_k$. Note that no similar

[8] The algebraic product, as functions of P_k, of the 'most informed' (13) and the 'least informed' (10) is equal to the product of the other two (11) and (12) arising from the intermediate levels of information

[9] Of course, it is likely that for some pairs $\langle p'_k, p''_k \rangle$ the two terms in curly braces may not be on the same side of unity, so that, for such pairs, when both sources of data are observed, a kind of cancellation will occur between the tendency of each kind of data separately to cause us to prefer p'_k over p''_k, or vice versa.

proportionate relationship holds when we consider updated *reliability predictions*, rather than updated distributions of the per demand failure probability P_k of the current sequence $\mathcal{A}_k$. Below, on p14, we compare the effects of these same four 'states of observation' on explicit *reliability predictions* (equations (29–32)). Nor is it possible to further factorise the right-most term in equations (14,15) since under our model we do not have the required conditional independence, given the value of P_k for the *current* observation sequence $\mathcal{A}_k$, of observations on *distinct previous* $\mathcal{A}_i$, $i < k$.

We remark that the approach used to obtain (13) is not limited to providing us with the updated univariate distribution of a single sequence's failure probability. The updated joint distribution of say $\langle P_{k-1}, P_k \rangle$ can be obtained from a bivariate form of the arguments. Since (11) and (12) are actually just special cases of (13) for certain of the n_i set equal to zero, we will not go through the extension to the bivariate case separately for each of the three observation cases distinguished in (11-13). In the general case where the observations take the form $\langle n_i, r_i \rangle_{i=1}^{k}$, which includes the three cases previously distinguished, we obtain an updated joint distribution

$$(P_{k-1}, P_k) \,\big|\, \langle n_i, r_i \rangle_{i=1}^{k} \sim \frac{p_{k-1}^{r_{k-1}}(1-p_{k-1})^{n_{k-1}-r_{k-1}} p_k^{r_k}(1-p_k)^{n_k-r_k} \int_{\theta \in \mathcal{S}} f_p(p_{k-1}|\theta) f_p(p_k|\theta) \big[\prod_{i=1}^{k-2} \int_0^1 p^{r_i}(1-p)^{n_i-r_i} f_p(p|\theta)\, dp\big] \mathrm{Prior}_\theta(\theta)\, d\theta}{\int_{\theta \in \mathcal{S}} \big[\prod_{i=1}^{k} \int_0^1 p^{r_i}(1-p)^{n_i-r_i} f_p(p|\theta)\, dp\big] \mathrm{Prior}_\theta(\theta)\, d\theta} \qquad (17)$$

from which we can, if we wish, investigate the sign and magnitude of any correlation between P_{k-1} and P_k (or powers of these) conditional on the observed data. For higher dimensional joint posterior distributions of the $\langle P_i \rangle$, (17) extends in the way you would expect, to give for the k-dimensional joint distribution

$$\langle P_i \rangle_{i=1}^{k} \,\big|\, \langle n_i, r_i \rangle_{i=1}^{k} \sim \frac{\Big\{\prod_{i=1}^{k} p_i^{r_i}(1-p_i)^{n_i-r_i}\Big\} \int_{\theta \in \mathcal{S}} \Big[\prod_{i=1}^{k} f_p(p_i|\theta)\Big] \mathrm{Prior}_\theta(\theta)\, d\theta}{\int_{\theta \in \mathcal{S}} \Big[\prod_{i=1}^{k} \int_0^1 p^{r_i}(1-p)^{n_i-r_i} f_p(p|\theta)\, dp\Big] \mathrm{Prior}_\theta(\theta)\, d\theta}$$

We will not pursue this investigation further in this paper and will concentrate on the updated distribution of the univariate P_k, and its consequences for reliability predictions of the single sequence $\mathcal{A}_k$ given the various combinations of observations discussed above.

Depending on the choice of the distribution family $\{f_p(\cdot|\theta); \theta \in \mathcal{S}\}$ and of the distribution Prior_θ, we may anticipate some analytic and computational difficulties in obtaining these updated distributions for P_k. However, we can perhaps more easily obtain expressions for the effect of the learning on the *moments* of the distribution of P_k. In fact the moments of these three alternative updated P_k-distributions (which will play the role of mixing distribution in Bayesian reliability prediction for $\mathcal{A}_k$) are important since any probability prediction of future failures of $\mathcal{A}_k$ is equivalent to the expectation, with respect to one of these updated P_k-distributions, of the equivalent prediction conditioned on P_k; and the latter conditional probability will generally involve *positive integer powers of* P_k. (See e.g. (1).) This follows because our model assumptions tell us that the three quantities *past failure behaviour of* $\mathcal{A}_k$, *future failure behaviour of* $\mathcal{A}_k$, *failure behaviour of other sequences* are conditionally independent given P_k. For example, the predictive probability of $R{=}r$ failures in n further demands on $\mathcal{A}_k$ is obtained by substituting the appropriate one of (11), (12), or (13) for $f_p(p|\theta)$ in (3).

More generally if:–

1. the term *observations* refers to some partial or complete joint observation of past failure behaviours of $\mathcal{A}_k$ and of other sequences $\mathcal{A}_i$; and
2. the term *future failure behaviour of* $\mathcal{A}_k$ refers to some pre-specified *event* concerning the pattern of future failure of sequence $\mathcal{A}_k$;

then we have

$$\mathbf{P}(\textit{future failure behaviour of } \mathcal{A}_k \,|\, \textit{observations}) = \\ \mathbf{E}\big(\, \mathbf{P}(\textit{future failure behaviour of} \mathcal{A}_k \,|\, P_k) \;\big|\; \textit{observations} \,\big) \quad (18)$$

where, on the right-hand side, the value of the inner probability will be a function of P_k (calculated as for an ordinary Bernoulli trials process) and where the outer expectation is *calculated with respect to the updated distribution of* P_k given '*observations*', which distribution will be one of equations (11), (12), and (13) when '*observations*' is of one of the three specific kinds we have discussed explicitly.

For an alternative perspective on the same predictions we remark that we are not *obliged* to think of them in terms of the updated distributions (11–13) of P_k. We can instead use the doubly stochastic structure of our model and its two layers of conditional independence[10] assumptions to show that a prediction of the form (18) will in fact assume a ratio form which can be understood directly in terms of two layers of nesting of probabilities and expectations with respect to our initial model distributions. In fact, our independence assumptions tell us that whenever '*observations*' is of such a 'product' form that we can decompose it into $\bigwedge_{i=1}^{k}(\textit{past behaviour of } \mathcal{A}_i)$ (i.e. if it is actually a conjunction of separate events concerning each $\mathcal{A}_i$ in isolation) then (18) can be shown to be equivalent to the formula

$$\begin{aligned} &\mathbf{P}(\textit{future failure behaviour of } \mathcal{A}_k \,|\, \textit{observations}) \\ &= \frac{\mathbf{E}\Big(\mathbf{P}(\textit{future and past failure behaviour of} \mathcal{A}_k \,|\, \theta) \prod_{i=1}^{k-1} \mathbf{P}(\textit{past failure behaviour of} \mathcal{A}_i \,|\, \theta)\Big)}{\mathbf{E}\Big(\prod_{i=1}^{k} \mathbf{P}(\textit{past failure behaviour of} \mathcal{A}_i \,|\, \theta)\Big)} \\ &= \frac{\mathbf{E}\Big(\mathbf{E}\big[\mathbf{P}(\textit{future f. b. of } \mathcal{A}_k \,|\, P_k)\, \mathbf{P}(\textit{past f. b. of } \mathcal{A}_k \,|\, P_k) \,\big|\, \theta\big] \prod_{i=1}^{k-1} \mathbf{E}\big[\mathbf{P}(\textit{past f. b. of } \mathcal{A}_i \,|\, P_i) \,\big|\, \theta\big]\Big)}{\mathbf{E}\Big(\prod_{i=1}^{k} \mathbf{E}\big[\mathbf{P}(\textit{past f. b. of } \mathcal{A}_i \,|\, P_i) \,\big|\, \theta\big]\Big)} \end{aligned} \quad (19)$$

where this last form is a prediction expressed as a ratio directly in terms of the distributions used to construct the model. In both numerator and denominator the inner probabilities are calculated as for a Bernoulli trials process, the inner expectations are obtained using the distribution $f_p(\,\cdot\,|\theta)$, and the outer expectations are taken with respect to our prior distribution Prior_θ.

Before making any observations, P_k has a marginal distribution whose m^{th} non-central moment is given by

$$\mathbf{E}(P_k^m) = \int_{\theta\in\mathcal{S}} \int_0^1 p^m f_p(p|\theta)\, dp\, \mathrm{Prior}_\theta(\theta)\, d\theta = \int_{\theta\in\mathcal{S}} \mathbf{E}(P_k^m \,|\theta)\; \mathrm{Prior}_\theta(\theta)\, d\theta \quad (20)$$

This moment of P_k is updated, by our three distinguished observation assumptions, to give expressions for the moments of the distributions (11), (12), and (13) which take the general form of ratios of expectations with respect to Prior_θ of multinomials in the moments of f_p (which moments are of course functions of Θ). This is a consequence of the fact that (11), (12), and (13) are simple linear transforms of this θ-parameterised p.d.f. of our assumed conditional distribution for P_k given Θ. (Or it can also be explained as a particular case of (19).) Specifically, taking the three observation cases in the same order as earlier, the m^{th} *updated* non-central moment of P_k is

$$\mathbf{E}(P_k^m \,|\, n_k, r_k) = \frac{\displaystyle\int_{\theta\in\mathcal{S}} \mathbf{E}\Big(P^{m+r_k}\,(1-P)^{n_k-r_k} \,\Big|\theta\Big)\; \mathrm{Prior}_\theta(\theta)\, d\theta}{\displaystyle\int_{\theta\in\mathcal{S}} \mathbf{E}\Big(P^{r_k}\,(1-P)^{n_k-r_k} \,\Big|\theta\Big)\; \mathrm{Prior}_\theta(\theta)\, d\theta}\,, \quad (21)$$

[10] of $\langle P_i \rangle$ given θ for the sequence family, and of success/failure on separate demands given P_i for a particular sequence $\mathcal{A}_i$

or

$$\mathbf{E}\Big(P_k^m\,\Big|\langle n_i,r_i\rangle_{i=1}^{k-1}\Big)=\frac{\displaystyle\int_{\theta\in\mathcal{S}}\mathbf{E}(P^m|\theta)\left[\prod_{i=1}^{k-1}\mathbf{E}\Big(P^{r_i}\,(1-P)^{n_i-r_i}\,\Big|\theta\Big)\right]\,\mathrm{Prior}_\theta(\theta)\,d\theta}{\displaystyle\int_{\theta\in\mathcal{S}}\left[\prod_{i=1}^{k-1}\mathbf{E}\Big(P^{r_i}\,(1-P)^{n_i-r_i}\,\Big|\theta\Big)\right]\mathrm{Prior}_\theta(\theta)\,d\theta}\,, \tag{22}$$

or

$$\mathbf{E}\Big(P_k^m\,\Big|\langle n_i,r_i\rangle_{i=1}^{k}\Big)=$$

$$\frac{\displaystyle\int_{\theta\in\mathcal{S}}\mathbf{E}\Big(P^{m+r_k}\,(1-P)^{n_k-r_k}\,\Big|\theta\Big)\left[\prod_{i=1}^{k-1}\mathbf{E}\Big(P^{r_i}\,(1-P)^{n_i-r_i}\,\Big|\theta\Big)\right]\,\mathrm{Prior}_\theta(\theta)\,d\theta}{\displaystyle\int_{\theta\in\mathcal{S}}\left[\prod_{i=1}^{k}\mathbf{E}\Big(P^{r_i}\,(1-P)^{n_i-r_i}\,\Big|\theta\Big)\right]\mathrm{Prior}_\theta\,(\theta)\,d\theta}\,, \tag{23}$$

respectively, under the three different assumptions: observation of $\mathcal{A}_k$ only; observation only of other sequences $\langle\mathcal{A}_i\rangle_{i=1}^{k-1}$; or observation of all of $\langle\mathcal{A}_i\rangle_{i=1}^{k}$. Note that here the left hand sides are updated expectations conditioned on observed data: The right-hand sides are *ratios* of *unconditional* expectations taken with respect to the original, prior Θ-distribution Prior_θ. The random variables whose unconditional expectations form these ratios are 'binomial-like' expressions in the moments of the distribution $f_p(\cdot|\theta)$, which, being deterministic functions of Θ, inherit their distributions from our chosen Prior_θ distribution. To emphasise this role played by these moments of P_k given θ, and at the same time to shorten equations (20–23) slightly, if we define

$$\mu_{r,s}(\theta)=\int_0^1 p^r(1-p)^s f_p\,(p|\theta)\,dp, \tag{24}$$

then we can write

$$\mathbf{E}(P_k^m)=\mathbf{E}(\mu_{m,0}) \tag{25}$$

$$\mathbf{E}(P_k^m\,|n_k,r_k)=\frac{\mathbf{E}(\mu_{m+r_k,n_k-r_k})}{\mathbf{E}(\mu_{r_k,n_k-r_k})}\,, \tag{26}$$

$$\mathbf{E}\Big(P_k^m\,\Big|\langle n_i,r_i\rangle_{i=1}^{k-1}\Big)=\frac{\mathbf{E}\left(\mu_{m,0}\displaystyle\prod_{i=1}^{k-1}\mu_{r_i,n_i-r_i}\right)}{\mathbf{E}\left(\displaystyle\prod_{i=1}^{k-1}\mu_{r_i,n_i-r_i}\right)}\,, \tag{27}$$

$$\mathbf{E}\Big(P_k^m\,\Big|\langle n_i,r_i\rangle_{i=1}^{k}\Big)=\frac{\mathbf{E}\left(\mu_{m+r_k,n_k-r_k}\displaystyle\prod_{i=1}^{k-1}\mu_{r_i,n_i-r_i}\right)}{\mathbf{E}\left(\displaystyle\prod_{i=1}^{k}\mu_{r_i,n_i-r_i}\right)}\,. \tag{28}$$

Equations (18) and (19) tell us that up-to-date reliability predictions may be similarly expressed as ratios of expectations of moments of $f_p(\cdot|\theta)$. Firstly, given no observation data at all, we have

$$\mathbf{P}(r_k|n_k)=\binom{n_k}{r_k}\mathbf{E}(\mu_{r_k,n_k-r_k}) \tag{29}$$

and, once having observed (only) that $R_k=r_k$, if r'_k is the number of failures predicted in a further n'_k demands on sequence $\mathcal{A}_k$,

$$\mathbf{P}(r'_k\,|n'_k,n_k,r_k)=\binom{n'_k}{r'_k}\frac{\mathbf{E}\Big(\mu_{r_k+r'_k,n_k+n'_k-r_k-r'_k}\Big)}{\mathbf{E}(\mu_{r_k,n_k-r_k})}\,, \tag{30}$$

For our other two observation assumptions we can write

$$\mathbf{P}\left(r_k \,\middle|\, n_k, \langle n_i, r_i\rangle_{i=1}^{k-1}\right) = \binom{n_k}{r_k} \frac{\mathbf{E}\left(\prod_{i=1}^{k} \mu_{r_i, n_i - r_i}\right)}{\mathbf{E}\left(\prod_{i=1}^{k-1} \mu_{r_i, n_i - r_i}\right)}, \tag{31}$$

$$\mathbf{P}\left(r'_k \,\middle|\, n'_k, \langle n_i, r_i\rangle_{i=1}^{k}\right) = \binom{n'_k}{r'_k} \frac{\mathbf{E}\left(\mu_{r_k + r'_k, n_k + n'_k - r_k - r'_k} \prod_{i=1}^{k-1} \mu_{r_i, n_i - r_i}\right)}{\mathbf{E}\left(\prod_{i=1}^{k} \mu_{r_i, n_i - r_i}\right)}. \tag{32}$$

(Note that the updated P_k-moments (25–28) are merely a special case of this prediction : the probability that the next m demands result in a string of m successive system failures.) In equation (29–32), the expectations occurring within the right-hand sides are taken with respect to the prior Prior_θ. So the conditioning observations are present in the right-hand expressions only through the specification of *which* moment-terms $\mu_{r,s}(\theta)$ comprise the products whose prior expectation is to be taken. Indeed, it may be useful to think of the distribution $f_p(\cdot|\theta)$, given a θ value, as represented by an infinite, 2-dimensional matrix of its moments $\mu_{r,s}(\theta)$. Then our choice of Prior_θ can be viewed as a distribution over these matrices. Our future reliability predictions will be expressed as product-expectations (over Θ) of certain elements from these matrices, where these elements are selected from the matrix at positions determined by the values of the failure counts we have observed in the past and by the precise future failure-count value whose predictive probability we wish to obtain.

4 An Upper Bound on Reliability Prediction : The Case of No Observed Failures

Consider the special case in which no failures at all have been observed—neither failures of the ⟨product, environment⟩ pair for which we specifically wish to predict reliability, nor failures of other pairs ⟨product, environment⟩ within the same family. This case may have importance as 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 of sequences within the family. Specialising the equations above to this case is simply a matter of substituting the observation $\langle r_i\rangle = \langle 0\rangle$. 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 ratios of expectations of products of the non-central moments[11] of $1-P$, with P coming from the distribution $f_p(\cdot|\theta)$. So, conclusions about the best reliability levels potentially measurable using this model can be thought of as dependent exclusively[12] on our decision about what may be considered realistic assumptions for our subjective prior distribution of the moment-vector

$$\langle \mu_{0,1}, \mu_{0,2}, \mu_{0,3}, \ldots\rangle = \left\langle \mathbf{E}(1-P\,|\theta)\,, \mathbf{E}\left((1-P)^2\,\middle|\theta\right), \mathbf{E}\left((1-P)^3\,\middle|\theta\right), \ldots\right\rangle \tag{33}$$

of the ⟨product, environment⟩ family.

Assuming that we do begin by believing that our family is highly reliable (to be more exact, that any individual ⟨product, environment⟩ pair within the family is highly likely to be highly

[11] i.e. moments of the probability of successful completion of an individual demand

[12] As far as reliability prediction is concerned, the significance of our specification and parameterisation $\{f_p(\cdot|\theta)\,;\,\theta\in\mathcal{S}\}$ of a collection of possible P-distributions, and the significance of our choice of prior Prior_θ over this collection, is contained entirely in the resulting distribution of the moment-vector (33).

reliable), then the conditional distribution of P given θ will be concentrated very close to 0 (for all except, perhaps, some values of the family parameter θ which we consider to be very unlikely, i.e. that are assigned small probability (density) values $\text{Prior}_\theta(\theta)$ by our prior for θ). Suppose a particular $f_p(\,\cdot\,|\theta)$, i.e. a particular value of the parameter θ, were highly reliable. This θ might correspond to say a particularly good design process, or perhaps a single product which is successful in achieving high operating reliability in a number of different operating environments. Then the first few at least of these moments $\mu_{0,i}$ ought to be very close to 1. But it now appears that it is the relative amounts by which we, at the outset, stochastically believe the higher moments are less than 1, and certain kinds of *correlations* in our beliefs about these moments (as functions of Θ) which determines how much our confidence in failure-free operation for $\mathcal{A}_k$ should grow when we observe failure-free operation of other sequences $\langle \mathcal{A}_i \rangle_{i=1}^{k-1}$. To understand precisely the sense in which the last statement is true substitute $\langle r_i \rangle = \langle 0 \rangle$ and $r'_k = 0$ in equations (29-32). This yields three expressions for the reliability function, i.e. the Bayesian predictive probability that the next m 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 $\langle \mathcal{A}_i \rangle_{i=1}^{k}$. These three alternative predictive probabilities of future consecutive successful demands on $\mathcal{A}_k$ should be compared with the unconditional

$$\mathbf{E}((1-P)^m) = \mathbf{E}(\mu_{0,m}) \ , \tag{34}$$

the probability that the next m demands on $\mathcal{A}_k$ will be failure-free given *no* conditioning observation of either $\mathcal{A}_k$ or any other sequences. Indeed it is the comparison of (29) with (31), and the comparison of (30) with (32) which indicate the impact of evidence from other sequences on our beliefs about the probability of failure-free operation, or *reliability function*, of sequence $\mathcal{A}_k$. In each case, the admission of evidence from sequences $\langle \mathcal{A}_i \rangle_{i=1}^{k-1}$ introduces a common factor $\prod_{i=1}^{k-1} \mu_{0,n_i}$ into the arguments of the unconditional $\mathbf{E}$-operators in both the numerator and the denominator.

5 Some Questions About Model Implications

Some particular questions of interest are:–

- How does our confidence in $\mathcal{A}_k$ behave as a function of the *number* of previous sequences, when these have all been observed to contain no failures for an equal number $n_i = n$ of demands?
- For a fixed number $k-1$ of previous sequences, observed for fixed periods, how much does one failure in one sequence spoil things as far as our confidence in sequence $\mathcal{A}_k$ is concerned? Then, how much does one failure in each of two of these sequences affect our conclusions? And so on, for $1, 2, 3, \ldots$ out of the $k-1$ previous sequences exhibiting one failure each, and the rest no failures?
- Is it best, given a fixed number, in total, of demands on previous $\langle \text{product}, \text{environment} \rangle$ pairs, to know that fewer $\langle \text{product}, \text{environment} \rangle$ pairs have shown failure-free operation over a larger number of demands each, or that a larger number of such pairs have each worked perfectly over a relatively small number of demands each? How important is this distinction, in terms of its effect on the size of the amount by which our confidence in $\mathcal{A}_k$ is improved by observation of the previous sequences?
- Where there have been some previous failures, and again keeping the total number of previous demands constant, do we prefer to hear that those failures have been concentrated amongst a small number, or even a single, previous sequence, or is it less depressing news for the current $\langle \text{product}, \text{environment} \rangle$ pair if we find that the previous sequences all showed a similar level of unreliability? (It seems obvious that, if we are especially interested in the reliability of the current sequence $\mathcal{A}_k$, then, given the choice, we should in general prefer observed failures to have been found in previous sequences, i.e. to be failures associated with software products or environments other than the current one.)

- Which, if any, of the answers to the above four groups of questions holds quite generally for all possible parametric distribution families $\{f_p(\cdot|\theta)\,;\,\theta{\in}\mathcal{S}\}$, and for all possible prior beliefs Prior_θ? Does a preference for, say, all failures to have occurred in a single previous sequence, rather than for the same total number of failures to have been distributed between several previous sequences, depend on specific characteristics of our assumed prior distributions?
- Extending further such consideration of the influence of our choice of Prior_θ, we might even ask, the family $\{f_p(\cdot|\theta)\,;\,\theta{\in}\mathcal{S}\}$ being assumed specified, about variation in the quantities of interest over *the space of all priors* Prior_θ, and, in particular, ask various, probably mathematically non-trivial, questions about extrema here. In practice Prior_θ ought ideally to capture genuine prior belief. However, given that the conclusions from this model are likely to be highly dependent on the shape of our prior belief, it is important to try to gain a general understanding of more precisely *how*, and to what extent, various different distributions Prior_θ will effect our conclusions. What are the extremes, in both the sense of extreme favourability and extreme unfavourability to high current reliability predictions, of the prior beliefs we might hold? Are the mathematical extremes here at all plausible in practice? Can we introduce geometric constraints on the shape of the prior distribution, such as unimodality, or continuous density function, or upper and lower limits on the values of P_i admitted as having positive probability, and how do such constraints effect the answers to our questions about extrema?

There are several other similar questions that can be asked, given our general model structure. We plan to address some of these issues in more detail in our on-going work. The next section contains some tentative results relating to some of these questions in the context of some simple instantiations of our general model.

Generally, in terms of the basic model structure of this paper, there is of course an added complication to these questions: It may well turn out that the questions as we have just listed them are insufficiently precise. What precisely does 'high current reliability' mean in the last bullet point above? It might transpire that the answers will depend on specifically how we choose to quantify the reliability of the present $\langle\text{product},\text{environment}\rangle$ pair. For example, in terms of the updated distribution (13) of P_k, or in terms of the associated[13] reliability function. And in each of these two cases, how do we compare two *functions*? Two alternative $\mathcal{A}_k$-reliability functions resulting from different observed behaviours of previous sequences might cross at future demand $m{=}10^4$, for example. That is to say, the previous-sequence observations which give the greater confidence in the current sequence's long-term reliability may give lower confidence in its short-term reliability. In such a case, the set of previous-sequence observations which we would prefer to see would depend on factors such as our predicted operational lifetime of the $\langle\text{product},\text{environment}\rangle$ pair $\mathcal{A}_k$.

6 Examples of Particular Choices of Prior Distributions for P given Θ, and for Θ

We shall retain throughout what follows our original assumptions that each sequence $\mathcal{A}_i$ constitutes a Bernoulli trials failure process with unknown parameter P_i, and that the $\langle P_i\rangle$ sequence is i.i.d. conditionally given an unknown sequence-family 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. To begin with, we will investigate a simple two-point distribution $f_p(\cdot|\theta)$ in §6.1. Though clearly a simplification, this model instantiation can be argued to have some practical relevance to attempts to certify 'ultra-high reliability $\langle\text{product},\text{environment}\rangle$s' as well as illustrating in a simple way the structure of our general model.

[13] using equation (18)

6.1 Two-point f_p, with θ interpreted as mass at fixed points of support, one of which is p=0

Suppose $P\,|\theta$ has a two-point distribution with θ equal to the probability[14] assigned to $p = 0$. So we assume

$$\mathbf{P}(P_i = 0\,|\theta) = \theta\,; \qquad \mathbf{P}(P_i = \pi\,|\theta) = 1 - \theta\,.$$

Thus we assume that, for each sequence $\mathcal{A}_i$ in the family, θ is the probability that $P_i = 0$. For example, we could imagine a formal verification technique is applied to each software product and that this technique fails—to deliver a perfect ($p = 0$) $\langle$product, environment$\rangle$—with an unknown probability $1-\theta$. When this happens, we assume that the resulting program failure probability is known; for example $\pi = 10^{-5}$ might be used, if these are high-integrity products. This assumption of a single known value for p whenever $p \neq 0$ would perhaps better be relaxed by allowing a distribution for p, but it simplifies the application of our general model, retains sufficient flexibility to provide a useful illustration of the model, and could perhaps be justified on the grounds of conservatism by assuming a worst case value π for the non-zero p. We can now apply our previous results to the analysis of this model. Though now containing a discrete distribution component, the model can if desired be obtained directly from the results of sections 2–3 by defining the common density of each of the P_i in terms of Dirac delta functions[15]

$$P\,|\theta \sim f_p(p|\theta) = \theta\,\delta(p) + (1-\theta)\,\delta(p-\pi), \tag{35}$$

say, where $0 < \pi < 1$ is fixed.

The likelihood of θ given periods $\langle n_i\rangle_{i=1}^k$ of observation of k sequences (c.f. equation (7)) is then

$$L(\theta\,;\langle n_i, r_i\rangle_{i=1}^k) = \prod_{1\le i\le k}\binom{n_i}{r_i}\cdot\prod_{\substack{1\le i\le k,\\ r_i>0}}\pi^{r_i}\,(1-\pi)^{n_i-r_i}\,(1-\theta)\cdot\prod_{\substack{1\le i\le k,\\ r_i=0}}\{(1-\pi)^{n_i}\,(1-\theta)+\theta\}$$

To within a factor which does not depend on θ we can write this as

$$L(\theta\,;\langle n_i, r_i\rangle_{i=1}^k) \propto L_k(\theta) = \prod_{\substack{1\le i\le k,\\ r_i>0}}(1-\theta)\cdot\prod_{\substack{1\le i\le k,\\ r_i=0}}\{(1-\pi)^{n_i}(1-\theta)+\theta\}$$

6.1.1 1st Case: General Prior$_\theta$

It follows that the posterior distribution of Θ given this data is now

$$\Theta\,\big|\langle n_i, r_i\rangle_{i=1}^k \sim \frac{L_k(\theta)\mathrm{Prior}_\theta(\theta)}{\int_{\theta\in\mathcal{S}} L_k(\theta)\mathrm{Prior}_\theta(\theta)\,d\theta}$$

In fact, since the parameter θ has a direct interpretation here as a probability, we must have $\mathcal{S} \subseteq [0,1]$, and we can assume without loss of generality that Prior$_\theta$ is extended in such a way that $\mathcal{S} = [0,1]$. We shall assume this has been done for the remainder of this section. If there has been a failure in the observed part of the current sequence (i.e. if $r_k > 0$), then the updated posterior distribution of P_k given our observation is trivially just $P_k = \pi$ with certainty. In this case, future reliability prediction is simply that of a Bernoulli trials failure process with parameter π. In the

[14] Contrast with the also interesting case where θ defines the *position* of the points of support of f_p—Or perhaps further generalisations where the positions and masses of two points of support for P are represented by a two or three dimensional θ.

[15] Provided we agree either to slightly extend our usual range $0 \le p \le 1$ of integration with respect to p, or to modify the usual definition of the δ-functions so that $\int_0^1 \delta(p)\,dp$ and $\int_0^1 \delta(p-1)\,dp$ should evaluate to 1 rather than $\frac{1}{2}$

interesting case where the current sequence $\mathcal{A}_k$ has so far exhibited no failure we have[16]

$$\mathbf{P}\Big(P_k = 0 \,\Big|\, \langle n_i, r_i\rangle_{i=1}^{k-1}, n_k, r_k{=}0\Big) = \frac{\int_0^1 \theta L_{k-1}(\theta)\mathrm{Prior}_\theta(\theta)\,d\theta}{\int_0^1 \{(1-\pi)^{n_k}(1-\theta)+\theta\}\, L_{k-1}(\theta)\mathrm{Prior}_\theta(\theta)\,d\theta}\,;$$

$$\mathbf{P}\Big(P_k = \pi \,\Big|\, \langle n_i, r_i\rangle_{i=1}^{k-1}, n_k, r_k{=}0\Big) = \frac{\int_0^1 (1-\pi)^{n_k}(1-\theta) L_{k-1}(\theta)\mathrm{Prior}_\theta(\theta)\,d\theta}{\int_0^1 \{(1-\pi)^{n_k}(1-\theta)+\theta\}\, L_{k-1}(\theta)\mathrm{Prior}_\theta(\theta)\,d\theta}\,.$$

Equation (18) tells us that this pair of probabilities may now be substituted in the following equation to obtain a reliability prediction for sequence $\mathcal{A}_k$

$$\mathbf{P}\Big(\text{No failure in next } m \text{ demands} \,\Big|\, \langle n_i, r_i\rangle_{i=1}^{k-1}, n_k, r_k{=}0\Big) = \mathbf{P}\Big(P_k = 0 \,\Big|\, \langle n_i, r_i\rangle_{i=1}^{k-1}, n_k, r_k{=}0\Big) + \mathbf{P}\Big(P_k = \pi \,\Big|\, \langle n_i, r_i\rangle_{i=1}^{k-1}, n_k, r_k{=}0\Big)(1-\pi)^m \tag{36}$$

This is equivalent to the $r_k = r'_k = 0, n'_k = m$ case of equation (32), where for this model structure we have

$$\mu_{r,s} = \begin{cases} (1-\pi)^s(1-\theta)+1, & \text{if } r = 0, \\ \pi^r(1-\pi)^s(1-\theta), & \text{if } r > 0. \end{cases} \tag{37}$$

Expressing the up-to-date distribution of P_k slightly more concisely using the 'odds' form mentioned in equations (14) and (15), we can compare it with the odds obtained from observation of only the present or only previous sequences. Cancelling some constants from the likelihood term $L_k(\theta)$ which occurs in both the numerator and the denominator, we find that

$$\frac{\mathbf{P}\Big(P_k = 0 \,\Big|\, \langle n_i\rangle_{i=1}^{k}, \langle r_i\rangle_{i=1}^{k} = \langle 0\rangle\Big)}{\mathbf{P}\Big(P_k = \pi \,\Big|\, \langle n_i\rangle_{i=1}^{k}, \langle r_i\rangle_{i=1}^{k} = \langle 0\rangle\Big)} = \frac{(1+y_k)}{y_k}\frac{\int_0^1 \prod_{i=1}^{k-1}(\theta+y_i)\,\theta\mathrm{Prior}_\theta(\theta)\,d\theta}{\int_0^1 \prod_{i=1}^{k-1}(\theta+y_i)\,(1-\theta)\mathrm{Prior}_\theta(\theta)\,d\theta} \tag{38}$$

$$\frac{\mathbf{P}\Big(P_k = 0 \,\Big|\, \langle n_i\rangle_{i=1}^{k-1}, \langle r_i\rangle_{i=1}^{k-1} = \langle 0\rangle\Big)}{\mathbf{P}\Big(P_k = \pi \,\Big|\, \langle n_i\rangle_{i=1}^{k-1}, \langle r_i\rangle_{i=1}^{k-1} = \langle 0\rangle\Big)} = \frac{\int_0^1 \prod_{i=1}^{k-1}(\theta+y_i)\,\theta\mathrm{Prior}_\theta(\theta)\,d\theta}{\int_0^1 \prod_{i=1}^{k-1}(\theta+y_i)\,(1-\theta)\mathrm{Prior}_\theta(\theta)\,d\theta} \tag{39}$$

$$\frac{\mathbf{P}(P_k = 0 \,|\, n_k, r_k{=}0)}{\mathbf{P}(P_k = \pi \,|\, n_k, r_k{=}0)} = \frac{(1+y_k)}{y_k}\frac{\mathbf{P}(P_k = 0)}{\mathbf{P}(P_k = \pi)}\,, \tag{40}$$

where the second term on the right-hand side of (40) is just the prior odds that $P_k = 0$ (before any portion of any sequence has been observed) and where we introduce the notation

$$y_i = \frac{1}{(1-\pi)^{-n_i} - 1}, \quad i = 1 \ldots k\,. \tag{41}$$

Note that for this two-point model with π assumed known, improving reliability estimates of the current sequence translate directly into an improving up-to-date probability of current $\langle$product, environment$\rangle$ perfection. This can be simply expressed in terms of the odds values given by equations (38)–(40). If the prior odds of $\mathcal{A}_k$-perfection is denoted o, and if equations (39) and (40) represent, respectively, improvements on this by factors of $\mathcal{R}$, achieved by means of previous $\mathcal{A}_i$-observation, and $\mathcal{R}'$, by means of direct observation of $\langle$product, environment$\rangle$ pair $\mathcal{A}_k$, then we will have a prior $\mathcal{A}_k$-perfection probability of $\big(1+\frac{1}{o}\big)^{-1} = \int_0^1 \theta\mathrm{Prior}_\theta(\theta)\,d\theta$, improving to posterior $\mathcal{A}_k$-perfection probabilities of $\big(1+\frac{1}{\mathcal{R}'o}\big)^{-1}$, $\big(1+\frac{1}{\mathcal{R}o}\big)^{-1}$, and, $\big(1+\frac{1}{\mathcal{R}\mathcal{R}'o}\big)^{-1}$, respectively, under the three different observation scenarios of equations (11–13). Clearly, it is the factor $\mathcal{R}$ which is of particular interest since it represents the advantage to be gained from incorporating data on the *previous* $\mathcal{A}_i$. There is a need to understand the way that this factor is determined by the combined effect of observations and of our prior distribution Prior_θ.

[16] This formula holds also for $k = 1$ if we put $L_0(\theta) = 1$

If failures in some previous sequences $\mathcal{A}_1 \ldots, \mathcal{A}_{k-1}$ *have* been observed then this is handled (irrespective of how many of these failures there were for each $\mathcal{A}_i$ which was seen to fail at least once) by replacing the corresponding factors $(\theta + y_i)$ by $(1-\theta)$ in both the numerator and denominator of equations (38) and (39). The expression in terms of y_i here brings to light an interesting limiting case

$$y_i \to \frac{1}{e^{n_i \pi} - 1} \tag{42}$$

which is likely to be a good approximation to reality for pairs $\langle \text{product}, \text{environment} \rangle$ which are very reliable, and which is obtained by letting $\pi \to 0$ and $n_i \to \infty$ (for those $\mathcal{A}_i$ which have been observed not to fail) whilst holding $n_i \pi$ constant for each of these sequences. In this limiting case, the updated distribution of P_k comes to depend only on the products $n_i \pi$, and not other than via these products on the values of n_i and π. Some idea of closeness to this limiting case can be obtained, in terms of the value of π, from the crude bounds

$$\frac{1}{e^{\frac{n_i \pi}{1-\pi}} - 1} < y_i < \frac{1}{e^{n_i \pi} - 1} \tag{43}$$

which are respectively obtained by applying the two well known inequalities

$$\left(1 + \frac{x}{n}\right)^n < e^x, \quad n > 0,\ x > 0 \qquad \text{and} \qquad \left(1 - \frac{x}{n}\right)^{-n} > e^x, \quad 0 < x < n. \tag{44}$$

For an interpretation of y_i we can say that y_i is a kind of inverse measure of the informativeness to us of our no-failures observation on sequence $\mathcal{A}_i$. Precisely, y_i is the odds of observing what has been observed (i.e. no failures) of sequence $\mathcal{A}_i$, under the assumption that its true failure probability is $P_i{=}\pi$. So $r_i{=}0$ with a y_i which is close to zero means that we have observed something about sequence $\mathcal{A}_i$ which would be extremely unlikely under the assumption $P_i{=}\pi$. Conversely, $r_i{=}0$ with a very large y_i means that, even if we somehow knew for certain[17] that $P_i{=}\pi$, we would still be virtually certain not to observe any failures of $\mathcal{A}_i$ in the n_i trials we have carried out. Equations (38–40) confirm that there is virtually no effect on our beliefs about P_k, arising from the observation that $r_i{=}0$, if the value of y_i is very large : In terms of inferences about P_k, large y_i makes the the observation $r_i{=}0$ almost equivalent to *no* observation of sequence $\mathcal{A}_i$ (i.e. almost equivalent to $n_i{=}0$)[18].

We developed our model of §§2–3 in a very general setting from which these two-point f_p results are a very special case. However, even for this simple model instantiation, several interesting and non-trivial questions can be asked about how much extra confidence in a current sequence $\mathcal{A}_k$ can be gained from the observation that previous pairs $\langle \text{product}, \text{environment} \rangle$ have performed well. The information that previous products have been observed to perform perfectly in their assigned environments over finite observation periods $\langle n_i \rangle_{i=1}^{k-1}$ is a special case of obvious interest for reasons stated earlier.

By concentrating first on this simple two-point case of our general model, we can avoid immediately having to grapple with many of the complications concerning alternative quantifications of reliability. For this two point model, it is effectively true to say that high reliability of the present $\langle \text{product}, \text{environment} \rangle$ $\mathcal{A}_k$ is unarguably equivalent to a large value of the updated probability $\mathbf{P}\left(P_k{=}0 \,\middle|\, \langle n_i, r_i \rangle_{i=1}^{k}\right)$. So in the case of this 2-point model there does exist a *single number* which can be said to represent the current $\mathcal{A}_k$ reliability prediction. Thus we can unambiguously order the reliability predictions which would result from two different sets of past failure observations. For instance, (36) shows us that, with a fixed numerical value for π specified by the model, we will not experience the complication of two reliability functions, produced by different past-sequence behaviours, which cross at some number m of demands into the future. We proceed to examine some of the questions of §5 for this simple model.

[17] the worst possible belief we can hold about $\mathcal{A}_i$, however much observing we do, under this two-point model

[18] For a logically consistent definition of y_i in the vacuous $n_i{=}0$ case, we might use $y_i{=}\infty$ on the understanding that we can then simply cancel from equations (38–40) all the infinite factors involving y_i

Firstly, we can easily see from (40) that observation of failure-free operation in sequence $\mathcal{A}_k$, itself, will improve the odds that this sequence has P_k=0. The odds in favour of P_k=0 increase by a factor $1+1/y_k$, or $(1-\pi)^{-n_k}$. (We note that this factor is not influenced by our prior beliefs about Θ.) This remains true, and by an identical proportion (see comments on p11), irrespective of whether observation of previous sequences has occurred, and irrespective of what failure behaviour was observed for those sequences. Intuitively we might expect that observing perfect failure-free behaviour in *previous* sequences should improve the odds that P_k=0 in a similar sort of consistent way, if not to the same extent. To confirm such an effect, and to investigate its magnitude we need to look at the ratio

$$\mathcal{R} = \frac{\mathbf{P}\left(P_k{=}0 \,\middle|\, \langle n_i\rangle_{i=1}^{k}, \langle r_i\rangle_{i=1}^{k} = \langle 0\rangle\right)}{\mathbf{P}\left(P_k{=}\pi \,\middle|\, \langle n_i\rangle_{i=1}^{k}, \langle r_i\rangle_{i=1}^{k} = \langle 0\rangle\right)} \Bigg/ \frac{\mathbf{P}(P_k{=}0 \,|\, n_k, r_k{=}0)}{\mathbf{P}(P_k{=}\pi \,|\, n_k, r_k{=}0)} \;=$$

$$\frac{\displaystyle\int_0^1 \prod_{i=1}^{k-1}(\theta+y_i)\,\theta\mathrm{Prior}_\theta(\theta)\,d\theta}{\displaystyle\int_0^1 \prod_{i=1}^{k-1}(\theta+y_i)\,(1-\theta)\mathrm{Prior}_\theta(\theta)\,d\theta} \Bigg/ \frac{\displaystyle\int_0^1 \theta\mathrm{Prior}_\theta(\theta)\,d\theta}{\displaystyle\int_0^1 (1-\theta)\mathrm{Prior}_\theta(\theta)\,d\theta}$$

To do this we will first introduce some slightly more concise versions of our existing notation and make some definitions. Define

$$l(\theta) = \prod_{i=1}^{k-1}(\theta+y_i)\,, \quad \theta_1 = \mathbf{P}(P_k{=}0) = \int_0^1 \theta\mathrm{Prior}_\theta(\theta)\,d\theta\,, \quad \text{and} \quad \theta_2 = l^{-1}\left(\int_0^1 l(\theta)\mathrm{Prior}_\theta(\theta)\,d\theta\right).$$

Note that we know from the convexity of l, using Jensen's inequality, and from l's monotonicity, that, for any Prior_θ, we must have $0 \le \theta_1 \le \theta_2 \le 1$. In fact we will have strict inequalities here except in some degenerate cases such as $y_i{=}\infty$, or where Prior_θ is a single point mass. With this notation it can be shown that the ratio representing our improvement of odds simplifies as follows

$$\mathcal{R} = \frac{\displaystyle\int_0^1 l(\theta)\theta\mathrm{Prior}_\theta(\theta)\,d\theta}{\displaystyle\int_0^1 l(\theta)(1-\theta)\mathrm{Prior}_\theta(\theta)\,d\theta}\,\frac{1-\theta_1}{\theta_1} = 1 + \frac{\displaystyle\int_0^1 l(\theta)\theta\mathrm{Prior}_\theta(\theta)\,d\theta - l(\theta_2)\theta_1}{\theta_1\displaystyle\int_0^1 l(\theta)(1-\theta)\mathrm{Prior}_\theta(\theta)\,d\theta} \tag{45}$$

where here the numerator and denominator of the ratio on the right-hand side are both positive and, together through this term, express the 'amount of benefit' obtained from observing the non-failure in the previous sequences. We know that the numerator of this ratio (i.e. of the amount by which the ratio $\mathcal{R}$ of the odds exceeds 1) on the right-hand side of (45) is non-negative because it expresses the *covariance*[19] of the random variables Θ and $l(\Theta)$: A random variable cannot be negatively correlated with any variable obtained by applying to it a non-decreasing function. In fact we can express this numerator as an integral of a non-negative function, in either of two slightly different ways

$$\begin{aligned}\int_0^1 l(\theta)\theta\mathrm{Prior}_\theta(\theta)\,d\theta - l(\theta_2)\theta_1 &= \int_0^1 \big(\theta-\theta_1\big)\big(l(\theta)-l(\theta_1)\big)\mathrm{Prior}_\theta(\theta)\,d\theta \\ &= \int_0^1 \big(\theta-\theta_2\big)\big(l(\theta)-l(\theta_2)\big)\mathrm{Prior}_\theta(\theta)\,d\theta\,. \end{aligned} \tag{46}$$

It is clear from (45) that the improvement in the odds that P_k=0 which results from the previous sequence observations can be thought of as the result of three interacting influences: the *original odds* (prior to observation of either this or any other sequence) captured in terms of the value of

[19] deriving from our assumed prior distribution Prior_θ

θ_1; the actual detailed description of the *observation of previous sequences* (both their number, and what is observed of each), which we can think of as being summarised by the function $l(\theta)$[20]; and, going beyond the simple *expectation* θ_1, the exact *shape* Prior_θ *of our prior beliefs about* θ. In terms firstly of the previous sequence observations, we can see that $l(\theta)$=constant, corresponding to a large y_i value for each sequence observed, is equivalent to a lack of useful information observed from the previous sequences. At the opposite extreme, the function $l(\theta) = \theta^{k-1}$ is the upper bound on the proportionate variability of l over the unit interval. This represents an upper bound on the improvement of our beliefs about a k^{th} sequence that can arise from observation of periods of perfect operation of the $k-1$ previous sequences.

Now, looking instead at the influence on $\mathcal{R}$ of the form of Prior_θ, we see that, for *whatever* set of past-sequence observations, (45) will have approximately the value 1 in the case where Prior_θ approaches the distribution of a degenerate, constant, random variable. I.e. if we are already more or less certain before observation commences, that $\Theta \approx \theta_1$, then one sequence will have little to tell us about another. At the other extreme of the form of Prior_θ for the same fixed mean θ_1, it seems that high levels of *variation* or *spread* in our prior subjective Θ-distribution will have the opposite effect, magnifying to its limits the significance for sequence $\mathcal{A}_k$ of what we have observed from previous sequences. For a particular function $l(\theta)$, these limits are finite, and so we might investigate them further. But we do this as a way of obtaining a slack upper bound on *how much* previous sequences could ever tell us (within the two-point f_p model of this section) rather than because we believe the extreme of Prior_θ-variance is likely to be a realistic model of a person's true prior beliefs about reliability variation within the family of software $\langle \text{product}, \text{environment} \rangle$ pairs. For fixed mean θ_1 the most extreme spread in prior beliefs about Θ is given by the distribution Prior_θ which consists of two point masses: θ_1 at Θ=1 and $1-\theta_1$ at Θ=0. It is easy to see that this Prior_θ[21], when substituted in the left-hand side of (45) gives the P_k=0 odds-increase of

$$\mathcal{R} = \frac{l(1)}{l(0)} = \prod_{i=1}^{k-1} \frac{1+y_i}{y_i} = (1-\pi)^{-\{\sum_{i=1}^{k-1} n_i\}} \,. \tag{47}$$

In considering the influence of the shape of Prior_θ on the usefulness of previous sequence observations, a point worth making about the form of the obtainable odds-improvement (45) is the following. If we consider the set of probability distributions Prior_θ on the unit interval *having some common fixed mean* θ_1, and if we hold fixed the previous-sequence observations (i.e. specify some fixed function $l(\theta)$), then, as we vary Prior_θ within this set (which, mathematically, is a convex set in a suitable vector space of real measures), we will find that the extrema of (45) must be attained somewhere on the boundary of this set of distributions. This is because $\mathcal{R}$, regarded as a function of the distribution Prior_θ, has a monotonicity property along 'straight lines' in the set of candidate Prior_θ distributions : When Prior_θ is a mixture[22], say,

$$\text{Prior}_\theta(\theta) = \lambda p_1(\theta) + (1-\lambda) p_2(\theta) \,, \quad 0 < \lambda < 1 \,, \tag{48}$$

of two probability distributions on the unit interval, having a common mean θ_1, but different values, say $\mathcal{R}(p_1) < \mathcal{R}(p_2)$, of the ratio (45), then the value $\mathcal{R}(\text{Prior}_\theta)$ of (45) corresponding to the mixture will satisfy $\mathcal{R}(p_1) < \mathcal{R}(\text{Prior}_\theta) < \mathcal{R}(p_2)$. This in turn follows from the fact that the numerator[23] and denominator of (45) are both non-negative-valued linear functionals of the probability distribution Prior_θ[24]. This kind of reasoning can be used to confirm that (47) is indeed

[20] although, as we have already mentioned, it is the vector of products $n_i \pi$, or to be more exact, in the case where π is not very small, of y_i values defined by (41), which contains the significant part of the previous $\langle \text{product}, \text{environment} \rangle$ pairs' influence on our beliefs here, i.e. π as well as the n_i determine l

[21] strictly speaking a weighted sum of two Dirac delta functions, but note footnote 15 on p18

[22] the argument extends easily to more general mixtures than the discrete mixture of just two distributions used here

[23] use either the left-hand side of (45), or the right-hand side of (45) with the first form of the right-hand side of (46) used as the numerator

[24] essentially we are using the identity $\frac{\lambda a + (1-\lambda)A}{\lambda b + (1-\lambda)B} = \mu \frac{a}{b} + (1-\mu)\frac{A}{B}$ (a convex combination) for any pair of ratios $\frac{a}{b}$ and $\frac{A}{B}$ of positive numbers, where $\mu = \frac{\lambda b}{\lambda b + (1-\lambda)B}$.

the maximum possible value of $\mathcal{R}$ for a fixed observation function l and mean θ_1 (and in fact, as we see in (47), that the value of this maximum is actually the same for all $0 < \theta_1 < 1$). The value in (47) tends (using the same reasoning as for (43)) to a limit $\exp\left(\sum_{i=1}^{k-1} n_i \pi\right)$ as $\pi \to 0$ keeping each of the the product terms $n_i\pi$ in the exponent constant. In addition to this limiting value, we also have, for finite n_i, an upper bound $\exp\left(\sum_{i=1}^{k-1} n_i \frac{\pi}{1-\pi}\right)$ by the same reasoning as that used to produced (43).

It is difficult to imagine how a two-point Prior_θ, such as that required above to attain the maximum (47) effect of past sequences, could possibly arise in practice as a realistic model of subjective prior belief. Some further, more realistic restriction on the set of admissible shapes of the Prior_θ distributions which have some particular mean θ_1 is probably worth exploring. For example the question of how big (45) can become (for some fixed θ_1 and l) when we require that the prior distribution Prior_θ should be *unimodal* looks a more interesting one from a practical point of view, if, unfortunately, more difficult mathematically to solve.

6.1.2 2nd Case: Parametric Restriction of Prior_θ to Beta Family

Consider a further model specialisation in the form of the assumption that Prior_θ is a Beta distribution[25]. This assumption is convenient for numerical reasons since it allows us to expand out the θ-polynomials in (38) and integrate analytically, term by term, using

$$\int_0^1 \theta^m (1-\theta)^{m'} \text{Prior}_\theta(\theta)\, d\theta = \frac{\beta(a+m, b+m')}{\beta(a,b)} \tag{49}$$
$$= \frac{a}{(a+b)} \frac{a+1}{(a+b+1)} \cdots \frac{a+m-1}{(a+b+m-1)} \frac{b}{(a+b+m)} \frac{b+1}{(a+b+m+1)} \cdots \frac{b+m'-1}{(a+b+m+m'-1)}$$

Hence problems of optimisation (of say the ratio (45)) within this parametric restriction of the choice of prior beliefs Prior_θ become scalar optimisation problems with respect to the two independent variables a and b, rather than mathematically more difficult optimisations in which the independent variable is a 'point' lying in a convex set of general probability measures (contained within a larger vector space of general measures on the unit interval).

Note also that the Beta Prior_θ assumption contains, as the limiting cases $a, b \to 0$ with a/b constant, the largest-variance[26], 2-point Prior_θ distribution of a given mean, as mentioned above, and also contains, as the cases $a, b \to \infty$ with a/b constant, the degenerate Prior_θ under which there is no possibility to learn about P_k from observation of other $\mathcal{A}_i$, $i = 1 \ldots k-1$.

It is of interest[27] to fix the prior probability $\mathbf{P}(P_k{=}0)$ and to examine how variation of the parameters a, b of the Beta distribution for Θ, subject to such a constraint, affects the amount that can be learned from previous sequences. In fact, fixing this prior probability is equivalent simply to fixing the ratio a/b, i.e. the prior odds, say o, that $P_k{=}0$. If we reparameterise the Beta distribution in terms of this odds o and the parameter b, then, as we have just said, for fixed o, the greatest (or least) possibility, as measured by the ratio $\mathcal{R}$ of equation (45), for learning about the current $\langle \text{product}, \text{environment} \rangle$ pair from perfect behaviour of all observed previous $\mathcal{A}_i$, $1 \leq i < k$, corresponds to the extreme limiting case $b \to 0$ (or $b \to \infty$). In fact, for this model, we can verify that, as we might expect, in this situation where all previous sequences have shown perfect behaviour over their fixed observation periods, $\mathcal{R}$ is a monotonic non-increasing function of b. See Appendix A for the details of this proof. This $\mathcal{R}(b)$-monotonicity provides us some information, at least in the Beta case, about what happens as we vary the shape of the distribution Prior_θ, for fixed mean θ_1 (i.e. fixed o given by $o = \theta_1/(1-\theta_1)$), between the two extreme cases of the constant (zero variance) prior distribution corresponding to $\mathcal{R} = 1$ at one extreme, and the other extreme of

[25] Not to be confused with the assumption, in the example of the next sub-section, that θ *is* $\langle a, b \rangle$ the parameter-pair of a Beta distribution, where that Beta distribution is our f_p-distribution for P_i given θ.

[26] the coefficient of variation of the Beta(a,b) distribution is $\sqrt{b/\{a(a+b+1)\}}$

[27] because this will provide an upper bound on 'how much' a given amount of previous-sequence data can tell us, under this model : not because we wish to suggest that such an optimisation is a valid method of 'eliciting' the shape of genuine Bayesian prior beliefs, nor even that this bound will be close to approachable in a genuine analysis of real systems

the maximal-variance, 2-point distribution with mass only at Θ=0, 1 corresponding to the largest possible value of $\mathcal{R}$, given by (47). Fixing $o = a/b$ and varying b for our Beta Prior$_\theta$ is equivalent to moving along the line $b = oa$ in the $\langle a, b\rangle$-plane. In this plane of Beta distribution parameters, it is precisely the points *inside* the unit square (i.e. the pairs $\langle a, b\rangle$ with $\min(a, b) < 1$) which correspond to bimodal Beta distributions. The Beta distributions corresponding to the outside or boundary of this square (those for which $\min(a, b) \geq 1$) are unimodal Beta distributions. Hence, in the case of a Beta Prior$_\theta$, we can conclude that, moving back from infinity (the degenerate constant $\Theta = \frac{o}{o+1}$ prior) along the line $b = a/o$ towards the origin gives a steadily increasing variance of Prior$_\theta$, and simultaneously an increasing value of $\mathcal{R}$, with both the maximal variance, and the maximal $\mathcal{R}$, which can arise from a unimodal Beta prior, being attained on the boundary of the square at $\langle a, b\rangle = \langle 1, \frac{1}{o}\rangle$, if $o \geq 1$, or at $\langle a, b\rangle = \langle o, 1\rangle$, if $o \leq 1$. Expressions for the accompanying $\mathcal{R}$ values are obtained by substituting Prior$_\theta(\theta) \propto (1-\theta)^{o^{-1}-1}$ and Prior$_\theta(\theta) \propto \theta^{o-1}$ in the formula (45). If this movement towards the origin of the $\langle a, b\rangle$-plane is continued *inside* the unit square, then, as the origin is approached, increasingly extreme forms of bimodality in the prior for Θ result in the variance of Prior$_\theta$ approaching a maximum value of $o/(o+1)^2$, and $\mathcal{R}$ approaching the extreme limiting case of (47). This latter extreme case has the rather absurd interpretation of all sequences being known to have the same P_i value in advance of observation, but with uncertainty somehow persisting (despite such a strong belief in uniformity of failure rates) as to whether the actual value of this universal failure probability is 0 or π. It seems that this smaller $\mathcal{R}_{\max}$ arising from the restriction to unimodal priors might be a more realistic upper bound on the attainable size of the improvement $\mathcal{R}$. However, we have not answered the general question under a unimodal Prior$_\theta$ of how large $\mathcal{R}$ can be for given fixed observations periods throughout which the $k-1$ previous sequences have all been failure-free. We do not know how much greater we might be able to make $\mathcal{R}$ if we experiment with unimodal Prior$_\theta$ outside the Beta family. Mathematically this appears to be a difficult constrained optimisation problem.

We can also use the analytic tractability gained by this Beta restriction on Prior$_\theta$ to investigate the question[28] of how the shape of our prior beliefs affect the preferred allocation of a fixed number of demands between a number of past $\langle$product, environment$\rangle$ pairs $\mathcal{A}_i$, i=1,2,...,k-1. The algebra is a bit awkward, but even with this Beta family assumption for Prior$_\theta$, and while limiting ourselves to the more tractable cases of small k, we are still able to establish the following result: Our prior beliefs about the $\langle$product, environment$\rangle$ *perfection probability* $\theta = \mathbf{P}(P_i{=}0|\theta)$ are of sufficient importance that the answer to the question posed in the third bullet point at the beginning of §5 may be 'No', 'Yes', or something more complicated, depending on the *combined effects* of the shape Prior$_\theta$ of our prior beliefs and the total amount N (or Z, see below) of past product data we have available. This establishes a principle that there are qualitative, as well as quantitative, questions concerning what our model says about the influence of observations of past $\langle$product, environment$\rangle$ pairs which cannot be answered until we have described the shape of our prior uncertainty about reliability variation between the $\langle$product, environment$\rangle$ pairs of our family $\langle\mathcal{A}_i\rangle$. Suppose we have a total number N of demands to distribute between $k-1$ previous $\langle$product, environment$\rangle$ pairs $\mathcal{A}_1, \mathcal{A}_2, \ldots \mathcal{A}_{k-1}$ and that our objective is to increase our confidence in the reliability of $\mathcal{A}_k$ as much as possible. To simplify the notation slightly, we work in terms of $Z = N(-\log(1-\pi))$, which we might choose to think of as a quantification of the total amount of past-sequence observation 'adjusted' for the difficulty of our task of discriminating between P_k=0 and P_k=π. (Clearly, the closer π is to zero, the more difficult it becomes to discriminate, by means of data, between the two possibilities P_k=0 and P_k=π.) We can describe our allocation of this past data between the $k-1$ previous sequences by means of a vector $\langle\nu_1, \nu_2, \ldots, \nu_{k-1}\rangle$, with $0 \leq \nu_i \leq 1$, $\nu_1+\nu_2+\ldots+\nu_{k-1} = 1$ where $\nu_i = n_i/N$.

Taking first the simplest case of just two previous sequences, i.e. $k = 3$, equation (45) (using

[28] The third bullet point at the beginning of §5

(39) and (41)) can be written

$$\mathcal{R} = \frac{b}{a} \cdot \frac{\int_0^1 \left[e^{\nu_1 Z}\theta + (1-\theta)\right]\left[e^{\nu_2 Z}\theta + (1-\theta)\right]\theta^a(1-\theta)^{b-1}\,d\theta}{\int_0^1 \left[e^{\nu_1 Z}\theta + (1-\theta)\right]\left[e^{\nu_2 Z}\theta + (1-\theta)\right]\theta^{a-1}(1-\theta)^{b}\,d\theta} \tag{50}$$

Expanding the products of square-bracketed terms and using (49), this reduces to

$$\mathcal{R} = \frac{(a+2)(a+1)e^{(\nu_1+\nu_2)Z} + (a+1)b\left(e^{\nu_1 Z} + e^{\nu_2 Z}\right) + (b+1)b}{(a+1)ae^{(\nu_1+\nu_2)Z} + a(b+1)\left(e^{\nu_1 Z} + e^{\nu_2 Z}\right) + (b+2)(b+1)} \tag{51}$$

Remembering that $\nu_1 + \nu_2 = 1$ and taking $\left|\nu_1 - \frac{1}{2}\right|$ as our measure of unevenness of allocation of the N (or Z) past observations between $\mathcal{A}_1$ and $\mathcal{A}_2$, we find that

$$\begin{aligned}
\mathcal{R} &= \frac{(a+2)(a+1)e^{Z} + 2(a+1)be^{\frac{Z}{2}}\cosh\left[(\nu_1 - \frac{1}{2})Z\right] + (b+1)b}{(a+1)ae^{Z} + 2a(b+1)e^{\frac{Z}{2}}\cosh\left[(\nu_1 - \frac{1}{2})Z\right] + (b+2)(b+1)} \\
&= \frac{a+1}{b+1}\left[\frac{b}{a} + \frac{(a+b+2)\left(e^{Z} - \frac{b(b+1)}{a(a+1)}\right)}{(a+1)ae^{Z} + 2a(b+1)e^{\frac{Z}{2}}\cosh\left[(\nu_1 - \frac{1}{2})Z\right] + (b+2)(b+1)}\right]
\end{aligned} \tag{52}$$

and it is apparent that $\mathcal{R}$ is a monotonic function of $\left|\nu_1 - \frac{1}{2}\right|$ with bounds

$$\mathcal{R}_1 = \frac{(a+1)e^{Z} + b}{ae^{Z} + b + 1}, \qquad \text{achieved at } \nu_1 = 0, 1 \tag{53}$$

$$\mathcal{R}_2 = \frac{(a+2)(a+1)e^{Z} + 2(a+1)be^{\frac{Z}{2}} + (b+1)b}{(a+1)ae^{Z} + 2a(b+1)e^{\frac{Z}{2}} + (b+2)(b+1)}, \qquad \text{achieved at } \nu_1 = \tfrac{1}{2} \tag{54}$$

We showed earlier that, irrespective of how much direct observation of $\mathcal{A}_3$ has been done, $\mathcal{R}$ is the factor by which the odds that $\mathcal{A}_3$ is perfect are improved by the past sequence (in this case $\mathcal{A}_1$ and $\mathcal{A}_2$) observation. Thus we conclude, for this simple $k = 3$ case, that

- If our prior beliefs for θ are Beta(a, b) with $\frac{b(b+1)}{a(a+1)} = e^Z$, then our posterior probability that P_3=0 is unaffected by changes in allocation of a fixed total amount Z of past sequence observation between $\mathcal{A}_1$ and $\mathcal{A}_2$. In this case, the observation of $\mathcal{A}_1$ and $\mathcal{A}_2$ improves our odds that $\mathcal{A}_3$ is perfect by a fixed factor $\mathcal{R} = \frac{(a+1)b}{a(b+1)}$.
- If our prior beliefs are Beta(a, b) with $\frac{b(b+1)}{a(a+1)} > e^Z$ then our posterior probability that P_3=0 is a strictly increasing function of $\left|\nu_1 - \frac{1}{2}\right|$ (i.e. we prefer our previous observations to have been allocated as unevenly as possible between the two previous sequences $\mathcal{A}_1$ and $\mathcal{A}_2$). In this case we have $\mathcal{R}_2 \leq \mathcal{R} \leq \mathcal{R}_1$ as we vary ν_i. If all of these previous observations are concentrated on only one past $\mathcal{A}_i$, then the maximum possible improvement $\mathcal{R}$ of odds that P_3=0 is attained as $\mathcal{R}$=$\mathcal{R}_1$.
- If our prior beliefs are Beta(a, b) with $\frac{b(b+1)}{a(a+1)} < e^Z$ then our posterior probability that P_3=0 is a strictly decreasing function of $\left|\nu_1 - \frac{1}{2}\right|$ (i.e. we prefer our previous observations to have been allocated as evenly as possible between the two previous sequences). Here we have $\mathcal{R}_1 \leq \mathcal{R} \leq \mathcal{R}_2$ as we vary ν_i. If these previous observations are exactly evenly allocated between $\mathcal{A}_1$ and $\mathcal{A}_2$, then the maximum possible improvement $\mathcal{R}$ of odds that P_3=0 is attained as $\mathcal{R}$=$\mathcal{R}_2$. (Of course this is only possible to do *exactly* when N is even.)

So, in general, we have shown that, supposing N and π to be given (so that Z is fixed) then we cannot answer the question about whether a person prefers the observation of previous ⟨product, environment⟩ pairs to be allocated evenly between those previous $\mathcal{A}_i$ without first clarifying the shape of that person's prior beliefs about the unknown perfection probability parameter

θ. However we can draw a few conclusions of a more general nature for this k=3 case. Suppose that the person's prior probability that a randomly selected $\mathcal{A}_i$ is perfect $\mathbf{E}(\Theta)$ has been stated, and we know that their Prior$_\theta$ is in the Beta family. Then the ratio a/b is determined, and so, if $\mathbf{E}(\Theta) > \frac{1}{2}$, i.e. if $a > b$, it must be true for any value of Z that

$$\frac{b(b+1)}{a(a+1)} < 1 < e^Z$$

giving a preference for even allocation of observations between $\mathcal{A}_1$ and $\mathcal{A}_2$, whatever the exact values of a and b. Similarly it can be shown that if $\frac{1}{2} \geq \mathbf{E}(\Theta) > (1+e^{\frac{Z}{2}})^{-1}$ then the same preference will be found; whereas if $\mathbf{E}(\Theta) < (1+e^Z)^{-1}$ then the converse must apply and we will prefer the past observations to be concentrated as much as possible on a single $\mathcal{A}_i$. In terms of the stated prior expectation for Θ, the remaining possibility $< (1+e^Z)^{-1} < \mathbf{E}(\Theta) < (1+e^{\frac{Z}{2}})^{-1}$ corresponds to the situation in which the preference may be for or against even distribution of past observations between $\mathcal{A}_1$ and $\mathcal{A}_2$, depending on the exact values of the parameters a, b and the value of Z. These conclusions follow easily, with the restriction to a Beta Prior$_\theta$, from the facts that $\frac{b(b+1)}{a(a+1)}$ will lie between $\frac{a}{b}$ and $\frac{a^2}{b^2}$ for all $a > 0$, $b > 0$, and that the prior expectation is defined in terms of the Beta parameters by $\mathbf{E}(\Theta) = (1+\frac{b}{a})^{-1}$.

We can, without too much difficulty, gain some understanding of what happens when we are considering the effects of the allocation of a fixed amount of observations between *three* previous sequences $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$, i.e. in the case k=4. We take the Euclidean distance (which is proportional to the sample standard deviation of $\{\nu_1, \nu_2, \nu_3\}$)

$$r = \sqrt{\left(\nu_1 - \tfrac{1}{3}\right)^2 + \left(\nu_2 - \tfrac{1}{3}\right)^2 + \left(\nu_3 - \tfrac{1}{3}\right)^2} = \sqrt{\nu_1^2 + \nu_2^2 + \nu_3^2 - \tfrac{1}{3}}$$

between the points $\langle \nu_1, \nu_2, \nu_3 \rangle$ and $\left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle$ as a measure of *how unevenly the past observations are distributed* between the three available previous $\langle$product, environment$\rangle$ pairs $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$. The maximum allowable value of r is clearly $(1-\frac{1}{3})^2 + (\frac{1}{3})^2 + (\frac{1}{3})^2 = \sqrt{\frac{2}{3}}$. Equation (45), with the Beta(a,b) Prior$_\theta$, now expands to

$$\mathcal{R} = \frac{(a+3)(a+2)(a+1)e^Z + (a+2)(a+1)be^Z\left(e^{-\nu_1 Z} + e^{-\nu_2 Z} + e^{-\nu_3 Z}\right) + (a+1)(b+1)b\left(e^{\nu_1 Z} + e^{\nu_2 Z} + e^{\nu_3 Z}\right) + (b+2)(b+1)b}{(a+2)(a+1)ae^Z + (a+1)a(b+1)e^Z\left(e^{-\nu_1 Z} + e^{-\nu_2 Z} + e^{-\nu_3 Z}\right) + a(b+2)(b+1)\left(e^{\nu_1 Z} + e^{\nu_2 Z} + e^{\nu_3 Z}\right) + (b+3)(b+2)(b+1)} =$$

$$\frac{(a+3)(a+2)(a+1)e^Z + (a+2)(a+1)be^{\frac{2Z}{3}}\left(e^{(\frac{1}{3}-\nu_1)Z} + e^{(\frac{1}{3}-\nu_2)Z} + e^{(\frac{1}{3}-\nu_3)Z}\right) + (a+1)(b+1)be^{\frac{Z}{3}}\left(e^{(\nu_1-\frac{1}{3})Z} + e^{(\nu_2-\frac{1}{3})Z} + e^{(\nu_3-\frac{1}{3})Z}\right) + (b+2)(b+1)b}{(a+2)(a+1)ae^Z + (a+1)a(b+1)e^{\frac{2Z}{3}}\left(e^{(\frac{1}{3}-\nu_1)Z} + e^{(\frac{1}{3}-\nu_2)Z} + e^{(\frac{1}{3}-\nu_3)Z}\right) + a(b+2)(b+1)e^{\frac{Z}{3}}\left(e^{(\nu_1-\frac{1}{3})Z} + e^{(\nu_2-\frac{1}{3})Z} + e^{(\nu_3-\frac{1}{3})Z}\right) + (b+3)(b+2)(b+1)} \tag{55}$$

Note, for a, b fixed, and fixed $\langle \nu_i \rangle$, we have a limiting case, representing an upper bound on $\mathcal{R}$, of $\lim_{Z\to\infty} \mathcal{R} = (a+j)/a$, where j is the number of the ν_i that are non-zero. This limiting case corresponds to conclusive information that j of $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ are perfect, accompanied by a complete absence of operational observation on the other $3-j$ previous $\langle$product, environment$\rangle$ pairs. The expression (55) is more difficult to analyse as a function of $\langle \nu_1, \nu_2, \nu_3 \rangle$ (with $\nu_1 + \nu_2 + \nu_3 = 1$) than its one-dimensional counterpart (52) because of $\mathcal{R}$'s dependence on the *direction* as well as the modulus r of the 3-vector $\left\langle \nu_i - \frac{1}{3} \right\rangle$ of differences from the uniform allocation of observations between the previous three sequences $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$. However, as a first approach to understanding something of the behaviour of this expression we can try replacing both numerator and denominator[29] by low order Taylor expansions in the ν_i. For some argument values, such as Z sufficiently small[30], and the Beta parameters a and b lying within certain ranges, this results in an approximate form for $\mathcal{R}$ in a case where both the numerator and denominator depend much more on the *modulus* of $\left\langle \nu_1 - \frac{1}{3}, \nu_2 - \frac{1}{3}, \nu_2 - \frac{1}{3} \right\rangle$ than they do on its *direction*. In fact, we are able in each case (i.e.

[29] We note that the denominator is obtained from the numerator by replacing a, b by $a-1, b+1$ so the reasoning only has to be done once.

[30] Z will tend of necessity to be small for small π (which we hope π should be for highly reliable systems) since it becomes infeasible to carry out the very large number of demands then necessary to make Z larger

for both numerator and denominator) to contain the influence of the direction of this vector entirely within a remainder term which is negligible under these conditions on the parameters. Furthermore, in each case, this dependence on r can be approximated by a non-negative quadratic whose minimum is at r=0. This approximation, though somewhat simplistic and consequently restricted in terms of the range of parameter values for which it is accurate, is sufficient to guide us in the identification of some examples, analogous to the alternatives found in the k=3 case above, where the uniform allocation of observations to previous sequences, is either a global minimum, or a global maximum of $\mathcal{R}$, as desired. However, for this k=4 case, we can also show that for certain values a, b, Z, the two remainder terms which, only, are the terms influenced by the direction as well as the modulus r of the deviation $\langle \nu_1 - \frac{1}{3}, \nu_2 - \frac{1}{3}, \nu_2 - \frac{1}{3} \rangle$ from uniformity may become larger and acquire a significant role in determining $\mathcal{R}$'s behaviour. In some of these cases, in contrast to what we found above for the k=3 situation, the uniform allocation $\nu_i = \frac{1}{3}$, $i = 1, 2, 3$, may turn out to be *neither* the global minimum nor the global maximum point of $\mathcal{R}$.

Taking the two ν_i-dependent summands in the numerator obtained above, it is shown in Appendix B that if we define

$$\alpha = \frac{1}{2} \log \left(\frac{b+1}{a+2} \right) - \frac{Z}{6} \tag{56}$$

then a Taylor expansion of degree 2 at the point $\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \rangle$ in ν_i, $i = 1, 2, 3$ (still assumed confined to the plane $\nu_1 + \nu_2 + \nu_3 = 1$), with remainder term, gives us

$$\begin{aligned}
&(a+2)(a+1)be^{\frac{2Z}{3}} \sum_{i=1}^{3} e^{(\frac{1}{3} - \nu_i)Z} + (a+1)(b+1)be^{\frac{Z}{3}} \sum_{i=1}^{3} e^{(\nu_i - \frac{1}{3})Z} \\
&\quad = e^{\frac{Z}{2}} (a+1)b\sqrt{(a+2)(b+1)} \times \\
&\qquad\qquad \Big\{ 6\cosh(\alpha) + \cosh(\alpha) Z^2 r^2 + \frac{Z^3}{3} \sum_{i=1}^{3} \big(\nu_i - \frac{1}{3}\big)^3 \sinh\big[\alpha + u\big(\nu_i - \frac{1}{3}\big)Z\big] \Big\}
\end{aligned}$$

for some value $0 < u < 1$ (u probably varies with a, b, Z and the ν_i but is constrained to the unit interval). Of the terms inside the curly brackets, the final, remainder term is the only term that depends on the allocation proportions ν_i in a way not confined purely to a dependence on their modulus r. Hence, this term and a corresponding term (with different u) in the analogous Taylor expansion of the denominator of $\mathcal{R}$ are the two which when small enough to ignore, lead to a much simplified behaviour of $\mathcal{R}$. Note that this approximation (obtained by disregarding the remainder term) is best thought of as a quadratic in r—not rZ—because the coefficients are functions of Z, but not of r. We will not look in detail[31] at the general conditions affecting the size of the two remainder terms as proportions respectively of the numerator and denominator of $\mathcal{R}$. We merely clarify the situation a little by remarking that we can expand

$$\begin{aligned}
&\frac{Z^3}{3} \sum_{i=1}^{3} \big(\nu_i - \frac{1}{3}\big)^3 \sinh\big[\alpha + u\big(\nu_i - \frac{1}{3}\big)Z\big] = \\
&\qquad \frac{Z^3}{3} \Big\{ \sinh(\alpha) \sum_{i=1}^{3} \big(\nu_i - \frac{1}{3}\big)^3 \cosh\big[u\big(\nu_i - \frac{1}{3}\big)Z\big] + \cosh(\alpha) \sum_{i=1}^{3} \big(\nu_i - \frac{1}{3}\big)^3 \sinh\big[u\big(\nu_i - \frac{1}{3}\big)Z\big] \Big\},
\end{aligned}$$

that the three different arguments of the hyperbolic function in each sum all fall in an interval of length less than Z, and that, simply by virtue of the constraints $\nu_i \geq 0$, $\sum_{i=1}^{3} \nu_i = 1$ we can obtain by elementary calculus the constraints

$$-\frac{1}{36} \leq \max\Big(-\frac{r^3}{\sqrt{6}}, -\frac{1}{9} + \frac{r^2}{2}\Big) \leq \sum_{i=1}^{3} \big(\nu_i - \frac{1}{3}\big)^3 \leq \frac{r^3}{\sqrt{6}} \leq \frac{2}{9}.$$

[31] Lengthier analyses are possible, such a transformation of $\langle \nu_i - \frac{1}{3} \rangle$ to polar coordinates (r, ϕ) allowing a further expansion of both the numerator and the denominator of $\mathcal{R}$ into double series, each of whose terms is of the form $a_{ij} r^i \{ \begin{smallmatrix} \cos \\ \sin \end{smallmatrix} \}(j\phi)$ where ϕ is an angle describing the direction of the vector $\langle \nu_i - \frac{1}{3} \rangle$ in the plane $\sum_{i=1}^{3} \nu_i = 1$ (i.e. using a Fourier series expansion in terms of ϕ).

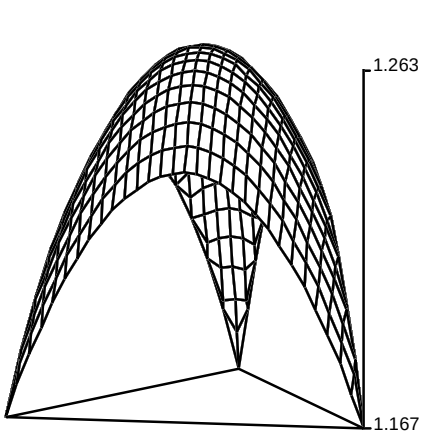
Thus, we have reduced $\mathcal{R}$ to an expression of the form

$$\mathcal{R} \approx K \cdot \frac{1 + C_1 r^2}{1 + C_2 r^2}, \qquad \text{with } K, C_1, C_2 > 0, \text{ and } 0 \leq r \leq \sqrt{\tfrac{2}{3}} \tag{57}$$

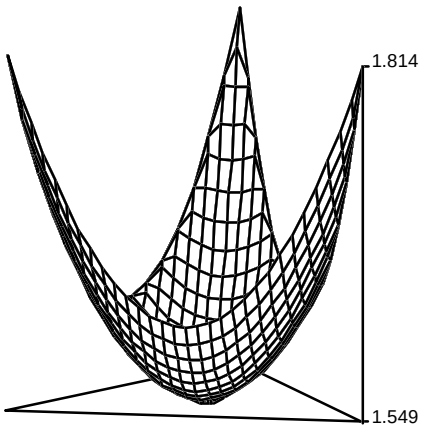
where $\langle K, C_1, C_2 \rangle$ can be thought of as a transform of the parameters $\langle a, b, Z \rangle$ of our model, but with the caveat that this approximation (57) is accurate (enough to serve as a useful model of the behaviour of the of the more complicated function $\mathcal{R}$ of (55)) only within some subdomain—defined by the requirement that the two neglected remainder terms should be sufficiently small—of the set of all possible values $a, b, Z > 0$. Then it is straightforward to conclude that if we hold a, b, Z fixed at some point within this subdomain, and vary the allocation vector $\langle \nu_i \rangle$, we will find $\mathcal{R}$ to be a monotonic function of r with r=0 being the global maximum, or minimum, respectively, as $C_1 < C_2$ or $C_1 > C_2$.

We have plotted $\mathcal{R}$ as a function of $\langle \nu_1, \nu_2, \nu_3 \rangle$ below for four different example values of $\langle a, b, Z \rangle$. Note that in each case it is the exact value (55) which is plotted. The approximation (57) was merely used as a guide to obtaining values of a, b, Z which achieve the three alternative general forms which $\mathcal{R}$ seems to display in this k=4 case, classified here as: decreasing from max at r=0 with approximate rotational symmetry[32]; increasing from min at r=0 with approximate rotational symmetry; and a third, catch-all category of 'other' more complex general behaviour.

If r is small and the parameters $\langle a, b, Z \rangle$ are within the right range, e.g. Z, $\alpha(a, b, Z)$ and $\alpha(a{-}1, b{+}1, Z)$ are not too large, then we have seen from (57) that $\mathcal{R}$ will be approximately an increasing function of r if $C_1 > C_2$, and a decreasing function if $C_1 < C_2$. Figure 2 shows two graphs illustrating this case. In these graphs, the base is the equilateral triangular surface:



Left plot : $\langle a, b, Z \rangle = \langle 5, 0.25, 2 \rangle$;



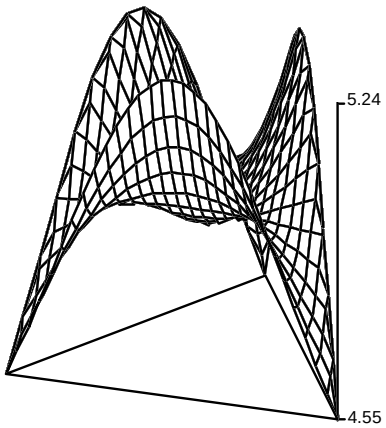
Right plot : $\langle a, b, Z \rangle = \langle 0.25, 5, 2 \rangle$

Figure 2: Plots of $\mathcal{R}$ vs. $\nu_1{+}\nu_2{+}\nu_3{=}1$

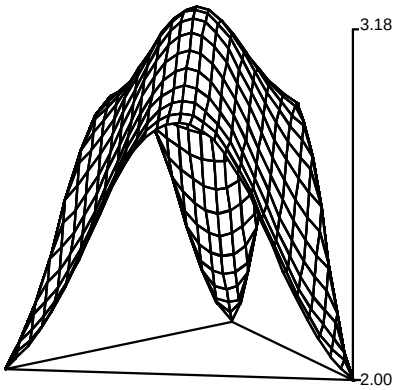
$\nu_1, \nu_2, \nu_3 > 0$, $\nu_1{+}\nu_2{+}\nu_3 = 1$. The vertical axis is the gain $\mathcal{R}$ of (55) obtained from observing that the three previous $\langle$product, environment$\rangle$ pairs have not failed. In the left plot we have $\langle C_1, C_2 \rangle = \langle .03, .16 \rangle$. In the right $\langle C_1, C_2 \rangle = \langle .38, .13 \rangle$.

Figure 3 shows examples where the situation is more complex with the remainder terms beginning to play a significant role so that we lose our approximate rotational symmetry of the plot. The left hand plot illustrates a case where the situation of even allocation among three previous $\langle$product, environment$\rangle$ pairs is intermediate (in terms of how much extra confidence it buys us in the current $\mathcal{A}_k$) to the cases of the same number of previous demands being either concentrated on a single previous $\langle$product, environment$\rangle$ pair, or being evenly allocated between two previous pairs. The right hand plot in Figure 3 is included to make the point that we are not suggesting that such odd behaviour of $\mathcal{R}$ will *necessarily* occur *everywhere* outside the domain of accuracy of our approximation (57). Here, the remainder terms are large but we still do have relatively

[32] i.e. approximately circular contours



Left plot : $\langle a, b, Z\rangle = \langle 0.25, 20, 5.3\rangle$;



Right plot : $\langle a, b, Z\rangle = \langle 1, 20, 10\rangle$

Figure 3: Plots of $\mathcal{R}$ vs. $\nu_1+\nu_2+\nu_3=1$

uncomplicated behaviour in the sense that there is a global maximum at $\langle \nu_i \rangle = \left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle$, with $\mathcal{R}$ a decreasing function of r.

It is worth underlining that it is often the *interaction* between prior beliefs and amount of past $\langle$product, environment$\rangle$ data that determines which of the cases illustrated in these plots applies; rather than either one of these things alone. These findings about the effects of the shape of a Beta Prior_θ on the preference for an even or an uneven allocation of previous sequence observations between 2 or 3 previous sequences raise two interesting questions which in this paper we have not explored:–

- What, if anything, can we say generally about the preferred allocation among larger numbers $k-1$ of previous sequences?
- To what extent are the precise results we found here arbitrary, accidental consequences of the fact that we happen to have restricted our priors to the *Beta* family? Perhaps some of these results are in fact particular cases of effects that could be stated in a framework of more general geometric constraints on the Prior_θ distribution without the need to constrain Prior_θ to a particular parametric family?

6.2 Use of a Beta family for f_p

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 to both the binomial and the negative binomial (including geometric) distributions and also provides a unique representation[33] of each possible[34] $\langle$mean,standard deviation$\rangle$ pair for a random variable P confined to the interval $[0, 1]$. If we use this as our f_p distribution family, we obtain a mixed process for the failures in each single sequence for which the probability of r failures in n demands is given by 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

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

[33] provided that limiting cases of the Beta parameters a, b are included

[34] i.e., all pairs in the closed half disk $\{\langle \mu, \sigma\rangle\ ;\ \mu, \sigma \geq 0 \wedge (\mu - \frac{1}{2})^2 + \sigma^2 \leq \frac{1}{4}\}$

in this case.

The likelihood (8) resulting from observation of k products operating in k allocated environments 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 an expression proportional to the likelihood of $\langle a, b\rangle$.

Having decided to investigate the Beta f_p, the choice of Prior_θ over $\mathcal{S}$, the positive quadrant[35], remains problematic. In real life there would be an 'expert' from whom we would wish to elicit the distribution that truly reflects his/her a priori belief. This is not an easy task in such a complex model, and the expert may find it difficult to represent his/her 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,

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

Firstly we can examine characteristics of the prior distribution (10) 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 our ignorance $\text{Prior}_\theta(a,b)$ (first expanding in partial fractions with respect to b in the case of the second moment) to give the first two cases of equation (20). But the centrally important effect of our model is to represent the effect of observed failure behaviour on both the distribution of P_k, and perhaps even more of interest, the reliability function, or probability of a future period of failure-free behaviour of a given length. The prior reliability function is given from equations (24) and (29) by

$$\begin{aligned}\mathbf{P}(X_k > n) = \mathbf{E}(\mu_{0,n}) &= \int_{a_1}^{a_2}\int_{b_1}^{b_2}\frac{\beta(a,b+n)}{\beta(a,b)}\frac{db\,da}{(a_2-a_1)(b_2-b_1)} \\ &= \int_{a_1}^{a_2}\int_{b_1}^{b_2}\frac{b(b+1)\ldots(b+n-1)}{(a+b)(a+b+1)\ldots(a+b+n-1)}\frac{db\,da}{(a_2-a_1)(b_2-b_1)}\,,\end{aligned}$$

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

Now to explore the effects of learning from observation we examine the realisations under these particular distributional assumptions of: firstly the posterior distributions for P_k given by equations (11–13); and secondly the predictions of X_k, the time to next failure of $\mathcal{A}_k$ using equations (30–32)[36]. In the most general case of arbitrary periods of observation of some finite number of previous sequences, 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 (a,b)-plane), where the integrands in the numerator and denominator are each equal to some product of terms of the form

$$\mu_{r,n-r}(a,b) = \mathbf{E}\big(P^r(1-P)^{n-r}\,\big|\,a,b\big) = \int_0^1 p^r(1-p)^{n-r}\frac{p^{a-1}(1-p)^{b-1}}{\beta(a,b)}\,dp = \frac{\beta(a+r,b+n-r)}{\beta(a,b)}$$

[35] 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 the all the limiting cases of the Beta family

[36] —given that we choose to concentrate on the no-failures case, for reasons of its interest as an upper bound on assurable reliabilities. In the case where past failures have been observed, we simply use the obvious analogues of (30–32), derived similarly from (11–13) and (18)

$$= \frac{a(a+1)\ldots(a+r-1)b(b+1)\ldots(b+n-r-1)}{(a+b)(a+b+1)\ldots\ldots\ldots\ldots\ldots\ldots(a+b+n-1)}$$

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 used with these sequences 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 found that from the numerical point of view, both the asymptotic form of the log-gamma function, and also the Euler-Maclaurin series for sums of form

$$\sum_{j=0}^{n-1} \log\left(1 - \frac{y}{x+j}\right) , \quad \text{where } 0 < y < x$$

were useful in approximating and bounding the integrals of these terms for large n. (See Appendix C for details.) For the sake of illustrating the algebraic form of the formulas, however, we give examples of the predictions of our model for hypothetical cases in which a very small number of observations have been seen. Suppose we wish to predict the probability that $\mathcal{A}_4$ will fail r times in its next 6 demands. In the absence of any knowledge of the past we obtain the distribution

$$R \sim \binom{6}{r} \int_{a_1}^{a_2}\int_{b_1}^{b_2} \frac{a(a+1)\ldots(a+r-1)b(b+1)\ldots(b+5-r)}{(a+b)(a+b+1)\ldots\ldots\ldots\ldots\ldots\ldots(a+b+5)} \frac{db\,da}{(a_2-a_1)(b_2-b_1)}$$

If we are now informed that $\mathcal{A}_4$ has in fact failed in the past 2 times out of 4, then our posterior distribution of P_4 is

$$P_4 \sim \frac{p_4^2(1-p_4)^2 \int_{a_1}^{a_2}\int_{b_1}^{b_2} \frac{p_4^{a-1}(1-p_4)^{b-1}}{\beta(a,b)} \frac{db\,da}{(a_2-a_1)(b_2-b_1)}}{\int_{a_1}^{a_2}\int_{b_1}^{b_2} \frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)} \frac{db\,da}{(a_2-a_1)(b_2-b_1)}}$$

and our updated distribution for the number of failures in the next 6 demands on $\mathcal{A}_4$ is

$$R \sim \binom{6}{r} \frac{\int_{a_1}^{a_2}\int_{b_1}^{b_2} \frac{a(a+1)\ldots(a+r+1)b(b+1)\ldots(b+7-r)}{(a+b)(a+b+1)\ldots\ldots\ldots\ldots\ldots\ldots(a+b+9)} \frac{db\,da}{(a_2-a_1)(b_2-b_1)}}{\int_{a_1}^{a_2}\int_{b_1}^{b_2} \frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)} \frac{db\,da}{(a_2-a_1)(b_2-b_1)}}$$

If we retract the information about the past 2 out of 4 failures of $\mathcal{A}_4$ (i.e., suppose it has not been seen), and instead suppose that pairs $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$, have been observed to fail 0 times out of 2, 2 times out of 3 and 1 time out of 4, respectively, then our posterior distribution of P_4 is

$$P_4 \sim \frac{\int_{a_1}^{a_2}\int_{b_1}^{b_2} \frac{p_4^{a-1}(1-p_4)^{b-1}}{\beta(a,b)} \left(\frac{b(b+1)}{(a+b)(a+b+1)}\right)\left(\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)}\right)\left(\frac{ab(b+1)(b+2)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}\right)\frac{db\,da}{(a_2-a_1)(b_2-b_1)}}{\int_{a_1}^{a_2}\int_{b_1}^{b_2} \left(\frac{b(b+1)}{(a+b)(a+b+1)}\right)\left(\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)}\right)\left(\frac{ab(b+1)(b+2)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}\right)\frac{db\,da}{(a_2-a_1)(b_2-b_1)}}$$

Now, the updated distribution for R in the next 6 demands on $\mathcal{A}_4$ is

$$R \sim \binom{6}{r} \times$$
$$\frac{\int_{a_1}^{a_2}\int_{b_1}^{b_2} \left(\frac{a(a+1)\ldots(a+r-1)b(b+1)\ldots(b+5-r)}{(a+b)(a+b+1)\ldots\ldots\ldots\ldots\ldots\ldots(a+b+5)}\right)\left(\frac{b(b+1)}{(a+b)(a+b+1)}\right)\left(\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)}\right)\left(\frac{ab(b+1)(b+2)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}\right)\frac{db\,da}{(a_2-a_1)(b_2-b_1)}}{\int_{a_1}^{a_2}\int_{b_1}^{b_2} \left(\frac{b(b+1)}{(a+b)(a+b+1)}\right)\left(\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)}\right)\left(\frac{ab(b+1)(b+2)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}\right)\frac{db\,da}{(a_2-a_1)(b_2-b_1)}}$$

If this information about $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$ is supplemented by the knowledge that $\mathcal{A}_4$ has failed 2 times out of 4 in the past, then the two corresponding updated distributions are

$$P_4 \sim$$
$$\frac{\int_{a_1}^{a_2}\int_{b_1}^{b_2} \frac{p_4^{a+1}(1-p_4)^{b+1}}{\beta(a,b)} \left(\frac{b(b+1)}{(a+b)(a+b+1)}\right)\left(\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)}\right)\left(\frac{ab(b+1)(b+2)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}\right)\frac{db\,da}{(a_2-a_1)(b_2-b_1)}}{\int_{a_1}^{a_2}\int_{b_1}^{b_2} \left(\frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}\right)\left(\frac{b(b+1)}{(a+b)(a+b+1)}\right)\left(\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)}\right)\left(\frac{ab(b+1)(b+2)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}\right)\frac{db\,da}{(a_2-a_1)(b_2-b_1)}}$$

and

$$R \sim \binom{6}{r} \times$$

$$\frac{\int_{a_1}^{a_2}\int_{b_1}^{b_2}\left(\frac{a(a+1)\ldots(a+r+1)b(b+1)\ldots(b+7-r)}{(a+b)(a+b+1)\ldots\ldots\ldots\ldots\ldots(a+b+9)}\right)\left(\frac{b(b+1)}{(a+b)(a+b+1)}\right)\left(\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)}\right)\left(\frac{ab(b+1)(b+2)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}\right)\frac{db\,da}{(a_2-a_1)(b_2-b_1)}}{\int_{a_1}^{a_2}\int_{b_1}^{b_2}\left(\frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}\right)\left(\frac{b(b+1)}{(a+b)(a+b+1)}\right)\left(\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)}\right)\left(\frac{ab(b+1)(b+2)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}\right)\frac{db\,da}{(a_2-a_1)(b_2-b_1)}}$$

The above example is intended to provide an illustration of the general form of the results for this Beta $f_p(\cdot\,|a,b)$ case with prior Prior_θ uniform on a rectangle. Table 1 shows some results that are more representative of what we might see when dealing with real safety-critical systems. These illustrative numerical results are based upon the observation of three previous sequences, each for a period of 10^7 demands without a single failure. In Table 1 we can see how various different

Region of Uniform Prior				Given no Data		Given no failure of this ⟨product, envir.⟩		Given no failure of previous 3 ⟨product, envir.⟩s		Given failure neither of this nor of previous 3 ⟨product, envir.⟩s	
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 $0.\text{XXXX} \times 10^{-n}$

Table 1: Effect on Reliability Predictions of Observation of Non-Failure of Previous ⟨product, environment⟩ pairs

assumptions for Prior_θ affect the strength of the inferences concerning a fourth sequence in the same family which can be drawn from this sort of evidence of high reliability of three previous, similar ⟨product, environment⟩ pairs.

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 sequences 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, environment⟩ pair $\mathcal{A}_4$ is affected by the type 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 sequence; given only evidence of failure-free working of earlier sequences; 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 sequences makes only a modest contribution to our confidence in the current sequence when compared with actual evidence of failure-free working during that sequence itself (compare columns 8 and 10). Thus when we only have evidence from the previous $\mathcal{A}_i$, $1 \leq i \leq 3$, 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 sequence[37].

The evidence from previous perfect working during the *same* sequence, however, is more informative. It allows us to be much more confident that this product will work perfectly in this environment 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 previous sequences may be useful in the case of safety-critical systems, especially as it is likely to come with little or no cost to developers. 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 $\langle$product, environment$\rangle$ survive 10^7 demands, and to .99 when we are told, in addition, that three other $\langle$product, environment$\rangle$ pairs have also survived 10^7 demands. Putting it another way, this evidence of survival in previous sequences 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 sequence.

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 the pair $\mathcal{A}_4$. 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.

7 Some General Remarks about Expressing Reliability

It is common to speak loosely of system reliability using terms such as: 'a 10 to the minus 5 system'; or 'a system with MTTF[38] 10^7'. Here the units implied would typically be *failures/demand* and *demands*, respectively[39]. Whilst it is of course possible to give a precise meaning to these terms, some care is required in comparing predictive distributions and rates based on a relatively sophisticated Bayesian model such as the one discussed here. We have taken the decision to extend an originally Bernoulli Trials process model with simple geometrically distributed time-to-failure distributions by including explicitly within the mathematics a probabilistic representation of our own subjective uncertainty about each Bernoulli trials parameter P_i. One of the consequences of this decision is that our subjective uncertainty about the P_i may never now remain static, so long as some form of observation is allowed to take place. Hence, as has been apparent in the formulae

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

[38] mean-time-to-next-failure

[39] or perhaps, in the continuous time case, *failures/hour* and *hours*, but we continue to concentrate for simplicity on the discrete *demand-count* time metric.

derived above (e.g. (30–32)), our predictive distributions of time to next failure are no longer restricted to the geometric family. Then if we retain the notion of geometric time to failure as a psychological standard of comparison, we must be careful when we speak of 'a 10 to the minus 5 system' to be clear whether we intend 'a system whose probability of failure on *the next* demand is identical to that of a geometric random variable with parameter 10^{-5}'; or 'a system whose MTTF is that of a geometric random variable with parameter 10^{-5}'; or perhaps 'a system whose median (or, say, upper 99.9 percentile) is that of a geometric random variable with parameter 10^{-5}'; etc.

It follows from (18) that the Bayesian predictive distribution of the process of future failures of a particular sequence, based on our model, will always take the form of a mixture of Bernoulli trials process distributions. Such mixture processes are *exchangeable*. (Conversely [3, p217] states that there are no other exchangeable, boolean valued, infinite random sequences than those obtained by mixing Bernoulli trials processes.) A few properties of these mixture processes were given earlier in equations (3–6), with $f_p(\cdot|\theta)$ playing the role of mixing distribution in these equations. Since these mixed processes form a more general class than the class of Bernoulli trials processes used as a model for each $\mathcal{A}_k$ *given its parameter* p_k, the theoretical possibility is introduced that it may also require some care to *compare* two different predictions which may emanate from our model (i.e. resulting from two different hypothetical findings from observation). How do we state unambiguously that one observation scheme gives rise to a prediction of 'higher program reliability' than another? If we were dealing with pure Bernoulli trials process predictions, then we would be able to say, for two predictions with parameters (i.e., per-demand failure probabilities) say π_1 and π_2, with $\pi_1 < \pi_2$, that prediction 1 predicts 'higher reliability' than prediction 2 in every possible sense: mean time to failure, median time to failure, failure rate, reliability function, etc. On the other hand for many exchangeable process predictions such as will be produced by a Bayesian analysis of our model, the mean time to failure does not exist (is infinite). Also we may well find that the median time to failure of prediction 1 may be greater than the median for prediction 2, whilst the order of say the 75%-iles could be reversed, i.e. the reliability functions obtained from two different observation schemes, such as two of equations (30–32), as functions of n, could conceivably intersect so that prediction 1 asserts better short term reliability than prediction 2 but the comparison might turn out to be reversed for longer term reliability predictions.

We might choose to compare predictions based on different observation assumptions in terms of their instantaneous 'reliability' measure, given (for the observation scenarios we have considered explicitly) by the $n = 1$ case of expressions (30– 29). However, in doing so we should bear in mind that such a measure does not necessarily tell us which prediction has the highest up-to-date 'mission survival probability' for a mission of a given length $n \neq 1$. It may be possible to overcome some of these difficulties by suitable restrictions upon the mathematical forms of Prior_θ and f_p, but these would need to be 'obviously reasonable' in their own right. Clearly it would be wrong for example to force an unreasonable (i.e. not believed) prior upon a human expert.

8 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 [4], 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 software product operating in a particular environment: evidence from testing of the $\langle \text{product}, \text{environment} \rangle$ pair itself and evidence from previous experience of 'similar' pairs.

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 environments or 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, in section 6.1.2 or section 6.2, 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 further exploration of the impact of different kinds of evidence upon the conclusions. For example, in our examples here we concentrated most of our attention on what is in many respects the most interesting case : that of complete perfection of operation of the previous sequences. This is the best news that it is possible to have, but it would be interesting to look more carefully at some cases where there have been failures in the earlier sequences.

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 [8].

Finally, all this modelling depends upon the reasonableness of notions of statistical 'similarity' between different demand sequences. In this we are merely making more formal the extremely informal claims that experts make when they argue that the failure behaviour of one demand sequence 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.

Acknowledgement

This work was supported by the ESPRIT DeVa project 20072, the DTI/EPSRC Safety Critical Systems Research Programme's DATUM Project. An early part of this work has appeared in the proceedings of the SAFECOMP 95 conference. The work has benefited considerably from numerous critical comments and suggested improvements by colleagues working on these projects and colleagues at the Centre for Software Reliability.

Appendices

A Proof that the Best Attainable Improvement $\mathcal{R}$ of the Odds that P_k=0 which is Obtained by Incorporation of Previous Demand-Sequence Data is, for Fixed Prior Θ-Mean, an Increasing Function of Our Prior Variance of Θ

This proof concerns the influence of the spread of our prior distribution for Θ on the improvement in $\mathbf{P}(P_k{=}0)$ which results from taking account of fixed observations of non-failure of previous sequences. See p23 for a more complete statement of the context. We will show this result specifically for the two-point P-distribution of §6.1 where, under our further restriction of §6.1.2 to a Beta assumption for Prior_θ, the result translates into the property that $\mathcal{R}$ of equation (45) is a decreasing function of the Beta parameter b, for $o = a/b$ fixed. So we work in terms of the reparameterised beta Prior_θ

$$\text{Prior}_\theta(\theta) = \frac{\theta^{ob-1}(1-\theta)^{b-1}}{\beta(ob,b)}, \qquad \frac{\theta_1}{1-\theta_1} = o \quad \text{(see p21)}. \tag{58}$$

Questions remain about the precise form and extent of any generalisations of the monotonicity result beyond this case. On expanding the polynomial $l(\theta) = \sum_{i=0}^{k-1} c_i\theta^i$, the left-hand side of (45) becomes

$$\mathcal{R} = \frac{\sum_{i=0}^{k-1} c_i\beta(ob+i+1,b)}{o\sum_{i=0}^{k-1} c_i\beta(ob+i,b+1)}\,. \tag{59}$$

Since we are assuming that the $k{-}1$ previous products have all performed perfectly in their respective environments over the numbers of trials observed for each, $l(\theta)$ remains defined as on p21 so the coefficients $\langle c_i\rangle$ are all non-negative. Now, holding the odds parameter o constant and differentiating $\mathcal{R}$ with respect to b, we can verify a non-increasing function $\mathcal{R}(b)$ for all $b > 0$ as follows.

$$\frac{\partial\mathcal{R}}{\partial b} =$$
$$\left\{\sum_{i=0}^{k-1} c_i\left[o\psi(ob{+}i{+}1) + \psi(b) - (o{+}1)\psi(ob{+}b{+}i{+}1)\right]\beta(ob{+}i{+}1,b)\sum_{j=0}^{k-1} c_j\,\beta(ob{+}j,b{+}1)\right.$$
$$\left. -\sum_{j=0}^{k-1} c_j\beta(ob{+}j{+}1,b)\sum_{i=0}^{k-1} c_i\left[o\psi(ob{+}i) + \psi(b{+}1) - (o{+}1)\psi(ob{+}b{+}i{+}1)\right]\beta(ob{+}i,b{+}1)\right\} \tag{60}$$
$$\Big/\, o\left\{\sum_{i=0}^{k-1} c_i\beta(ob{+}i,b{+}1)\right\}^2$$

where ψ is the digamma function (put $r = 0$ in (78)) and where we have used the relations

$$\frac{\partial\beta}{\partial a}(a,b) = \beta(a,b)[\psi(a) - \psi(a+b)], \qquad \frac{\partial\beta}{\partial b}(a,b) = \beta(a,b)[\psi(b) - \psi(a+b)]\,. \tag{61}$$

The two double sums of the numerator of (60) can be expanded out and their corresponding terms subtracted to express the numerator as a single double sum of the form

$$\begin{aligned}\text{Numerator} &= \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} t_{ij}c_ic_j \\ &= \sum_{i=0}^{k-1} t_{ii}c_i^2 + \sum_{i=1}^{k-1}\sum_{j<i}(t_{ij}+t_{ji})c_ic_j\,.\end{aligned} \tag{62}$$

We proceed to show that neither of the summands in the second form (62) can be greater than 0. In examining the terms of (62) it simplifies matters to first extract from t_{ij} a positive factor

$$s_{ij} = \frac{b}{(ob+b+i)(ob+b+j)}\beta(ob+i,b)\beta(ob+j,b) = \frac{\Gamma(ob+i)\Gamma(ob+j)\Gamma(b+1)\Gamma(b)}{\Gamma(ob+b+i+1)\Gamma(ob+b+j+1)} \quad (63)$$

which is symmetric in i,j. It can then be shown by selecting the relevant terms from the numerator of (60) and using the recurrence formula $\psi(z+1) = \psi(z) + 1/z$ that $t_{ii}/s_{ii} = -i/b \leq 0$. To do the same for the off-diagonal sums of pairs of terms from (62) is slightly more cumbersome. Selecting the relevant terms from (60) and simplifying in a similar way, we are left with (assuming without loss of generality that $i > j \geq 0$)

$$\frac{t_{ij}+t_{ji}}{s_{ij}} = (i-j)\Big\{o\big[\psi(ob+i) - \psi(ob+j)\big] - (o+1)\big[\psi(ob+b+i+1) - \psi(ob+b+j+1)\big]\Big\} - \frac{i+j}{b} \quad (64)$$

The problem here is the curly-bracketed term: ψ is monotonic increasing and the fact that $o+1 > o$ suggests that this term might be negative, finishing our task rather easily. But perhaps the ratio $(o+1)/o$ is insufficiently large to compensate for the fact that the function ψ is *concave*. To verify that the whole expression (64) cannot be positive we manipulate it as follows, beginning by expanding the two ψ-differences

$$\begin{aligned}
\frac{t_{ij}+t_{ji}}{s_{ij}} &= (i-j)\sum_{h=j}^{i-1}\left(\frac{o}{ob+h} - \frac{o+1}{(o+1)b+h+1}\right) - \frac{i+j}{b} \\
&= (i-j)\sum_{h=j}^{i}\left(\frac{o}{ob+h} - \frac{o+1}{(o+1)b+h}\right) + (i-j)\left(\frac{o+1}{(o+1)b+j} - \frac{o}{ob+i}\right) - \frac{i+j}{b} \\
&= (i-j)\sum_{h=j}^{i}\frac{-h}{(ob+h)\big((o+1)b+h\big)} - \frac{(i+j)o(o+1)b^2 + ij(i+j+2b+4ob)}{b(ob+i)\big((o+1)b+j\big)} \leq 0
\end{aligned}$$

We can therefore finally conclude that $\frac{\partial \mathcal{R}}{\partial b} \leq 0$, making $\mathcal{R}$ a monotonic non-increasing function of $b > 0$ for $o = a/b$ fixed. Examining the reasoning above, we see that $\mathcal{R}$ will almost always be a *strictly* decreasing function of b, the exceptional cases occurring only when we consider a few special or limiting conditions on the number $k-1$ of previous sequences, the beta parameters o, b, and the possibility of zero values for the coefficients $\langle c_i \rangle$ in the expansion of the polynomial $l(\theta)$. For $\mathcal{R}(b)$ to be constant over some b-interval would require *all* terms in the sum (62) to be zero inside that interval.

B Taylor Expansion of Numerator of Improvement $\mathcal{R}$ in Odds of $\mathcal{A}_k$-Perfection that Results From Observation of $\langle \mathcal{A}_1, \ldots, \mathcal{A}_{k-1} \rangle$

We wish to expand the two middle terms of the numerator of (55) on p26 as a Taylor series in powers of $\langle \nu_1, \nu_2, \nu_3 \rangle$, at the point $\left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle$. Our expansion is required to hold only within the plain $\nu_1 + \nu_2 + \nu_3 = 1$. It is clear from the development on p27 that what it remains to do, is to show that the Taylor expansion for the term

$$t = \sqrt{\frac{a+2}{b+1}}e^{\frac{Z}{6}}\sum_{i=1}^{3} e^{(\frac{1}{3}-\nu_i)Z} + \sqrt{\frac{b+1}{a+2}}e^{-\frac{Z}{6}}\sum_{i=1}^{3} e^{(\nu_i - \frac{1}{3})Z} \quad (65)$$

is given by the series

$$s = 6\cosh(\alpha) + \cosh(\alpha)Z^2r^2 + \frac{Z^3}{3}\sum_{i=1}^{3}\big(\nu_i - \frac{1}{3}\big)^3 \sinh\big[\alpha + u\big(\nu_i - \frac{1}{3}\big)Z\big] \quad (66)$$

By (56) we can write t as

$$t = e^{-\alpha}\sum_{i=1}^{3} e^{(\frac{1}{3}-\nu_i)Z} + e^{\alpha}\sum_{i=1}^{3} e^{(\nu_i-\frac{1}{3})Z} \quad (67)$$

$$= 2\sum_{i=1}^{3}\cosh\left[\alpha + (\nu_i - \frac{1}{3})Z\right] \quad (68)$$

$$= F(Z), \qquad \text{say.} \quad (69)$$

We can now use a Taylor expansion for the function F

$$F(Z) = F(0) + F'(0)Z + F''(0)\frac{Z^2}{2} + F^{(3)}(uZ)\frac{Z^3}{6}, \qquad \text{where } 0<u<1$$

with

$$F^{(n)}(Z) = \begin{cases} 2\sum_{i=1}^{3}(\nu_i-\frac{1}{3})^n \cosh[\alpha + (\nu_i-\frac{1}{3})Z], & n = 0,2,4,\ldots \\ 2\sum_{i=1}^{3}(\nu_i-\frac{1}{3})^n \sinh[\alpha + (\nu_i-\frac{1}{3})Z], & n = 1,3,5,\ldots \end{cases}$$

to deduce the result that $t = s$ for some $0<u<1$ (with u depending on a, b, Z and $\langle\nu_i\rangle$). The first order term is zero because $\sum_{i=1}^{3}(\nu_i-\frac{1}{3}) = \left[\sum_{i=1}^{3}\nu_i\right] - 1 = 0$

C Numerical Approximation to Very High Order Non-Central Moments of the Beta Distribution

We require for the purpose of plotting and numerical integration in §6.2 to have an efficient algorithm for calculating the expectation of $(1-P)^n$ for large n when P is distributed with a beta distribution with parameters a, b. Thus we require an algorithm to calculate[40]

$$\begin{aligned} \mu_{0,n}(a,b) &= \frac{\beta(a,b+n)}{\beta(a,b)} \\ &= \frac{\Gamma(b+n)\,\Gamma(a+b)}{\Gamma(b)\,\Gamma(a+b+n)} \\ &= \frac{b(b+1)\ldots(b+n-1)}{(a+b)(a+b+1)\ldots(a+b+n-1)}\,. \end{aligned}$$

Problems with overflow, long computation times, and loss of precision due to subtraction of very similar numbers were experienced when attempting to compute using standard beta, gamma and log-gamma library functions in the obvious ways directly from the forms above. To avoid these problems some bounds are obtained below by directly working with the specific function $\mu_{0,n}(a,b)$.

Firstly, we note that

$$\log\left(\mu_{0,n}(a,b)\right) = \sum_{i=0}^{n-1}\log\left(1 - \frac{a}{a+b+i}\right)$$

so we can apply the 'integral test' approximation to sums of any strictly decreasing function f

$$\int_0^n f(t)\,dt < \sum_{i=0}^{n-1} f(i) < f(0) - f(n) + \int_0^n f(t)\,dt\,,$$

[40] Although, for our purposes we are only interested in integer n, we note in passing that moments of non-integer order are perfectly well defined and that for the beta distribution we have the curious symmetry $\mu_{0,n}(a,b) = \mu_{0,a}(n,b)$, apparent from the Gamma-function representation here.

where $f(t) = -\log\left(1 - \frac{a}{a+b+t}\right)$, to give the interesting bounds

$$\mathcal{F}[t \mapsto (t-1)\log(t)] < \log\left(\mu_{0,n}(a,b)\right) < \mathcal{F}[t \mapsto t\log(t)] \tag{70}$$

where $\mathcal{F}$ (or, strictly, $\mathcal{F}_{a,b,n}$) is the linear functional given by

$$\mathcal{F}[g] = -g(a+b+n) + g(b+n) + g(a+b) - g(b)\,, \quad \text{for functions } g\,,$$

i.e., intuitively, $\mathcal{F}$ applies to its scalar function argument a difference operator the 'spacing' of whose differences is specified by a and n and the 'location' of application of which is specified by b. Note that, when $\mathcal{F}$'s argument g can be differentiated twice, the identity

$$\mathcal{F}[g] = -\int_0^a \int_0^n g''(t_1 + t_2 + b)\, dt_2\, dt_1 \tag{71}$$

will sometimes be used in what follows to demonstrate monotonicity of expressions which involve $\mathcal{F}$. Since we are interested in cases where n (and sometimes b also) are large compared to a, there are likely to be subtraction problems with numerical accuracy in calculating the bounds in (70) and some other bounds and approximations which also turn out to be defined by the application of $\mathcal{F}$ to some function. These can be solved by rearrangement and perhaps also Taylor series approximations. E.g. for the upper bound, if n is much larger than a we can use

$$\begin{aligned}
&-(a+b+n)\log(a+b+n) + (b+n)\log(b+n) + (a+b)\log(a+b) - b\log(b) = \\
&\qquad -a\log(a+b+n) - (b+n)\log\left(\frac{a+b+n}{b+n}\right) + (a+b)\log(a+b) - b\log(b)
\end{aligned}$$

—unless b is also much larger than a, in which case

$$\begin{aligned}
&-(a+b+n)\log(a+b+n) + (b+n)\log(b+n) + (a+b)\log(a+b) - b\log(b) = \\
&\qquad -a\log\left(\frac{a+b}{a+b+n}\right) \\
&\qquad +b\left[-\frac{1}{2}\left(\frac{a}{b}\right)^2 + \frac{1}{3}\left(\frac{a}{b}\right)^3 - \ldots\right] - (b+n)\left[-\frac{1}{2}\left(\frac{a}{b+n}\right)^2 + \frac{1}{3}\left(\frac{a}{b+n}\right)^3 - \ldots\right]
\end{aligned}$$

should produce an accurate answer. Once these rather minor subtraction problems have been tackled, the resulting bounds[41]

$$\frac{(b+n)^{(b+n-1)}(a+b)^{(a+b-1)}}{b^{(b-1)}(a+b+n)^{(a+b+n-1)}} < \mu_{0,n}(a,b) < \frac{(b+n)^{(b+n)}(a+b)^{(a+b)}}{b^b(a+b+n)^{(a+b+n)}} \tag{72}$$

on $\mu_{0,n}(a,b)$ can be used for many values of n and ranges of $\langle a,b\rangle$ to produce quite tight bounds on the reliability predictions discussed in §6.2. These bounds are themselves in the ratio

$$1 + \frac{an}{b(a+b+n)}$$

so that, for example, when a is small compared to b, we know that these bounds are at least correspondingly accurate approximations to $\mu_{0,n}(a,b)$. Returning to our application in §6.2, it is worth remarking that such values of a and b give very plausible distributions for P to characterise a family of $\langle$product, environment$\rangle$ pairs designed for very high reliability.

But, for those values, e.g. when $\frac{b}{a}$ is small, where these bounds are not known to give satisfactory accuracy, we can use the more general[42] and tighter bounds obtained using the Euler-Maclaurin summation formula, as follows.

[41] but note the improvement to the upper bound mentioned later on p41

[42] in that they are useful over a wider region of $\langle a,b,n\rangle$

Abramowitz and Stegun [1, p257] give bounds on the remainder $S_r(t)$ of the asymptotic expansion of the log-gamma function

$$\log(\Gamma(t)) = (t-\tfrac{1}{2})\log(t) - t + \tfrac{1}{2}\log(2\pi) + \sum_{k=1}^{r} \frac{B_{2k}}{2k(2k-1)} \frac{1}{t^{2k-1}} + S_r(t)\,. \tag{73}$$

Here, $B_0, B_1, B_2, \ldots$ are the Bernoulli numbers $1, -\frac{1}{2}, \frac{1}{6}, \ldots$, see [1, pp804–10]. Note that if the terms in negative powers of t here, and the remainder term $S_r(t)$ are all neglected, and the remaining part of the right-hand side of (73) is then substituted in the definition of $\log(\mu_{0,n}(a,b))$, we obtain the approximation $\mathcal{F}[t \mapsto (t-\frac{1}{2})\log(t)]$, where we have used the same linear functional notation $\mathcal{F}$ introduced above. It is straightforward to show[43] that this asymptotic approximation to $\log(\mu_{0,n}(a,b))$ lies between the two 'integration test' bounds obtained above. If, in place of (73), we consider instead the slightly easier problem of asymptotic approximation to *difference* of two values of the log-gamma function at arguments separated by an *integer*[44], then we can obtain information about the corresponding remainder term directly from the Euler-Maclaurin summation formula[45] [12, pp478–82]

$$\sum_{i=0}^{n-1} f(i) = \int_0^n f(t)\,dt + \frac{f(0)}{2} - \frac{f(n)}{2} + \sum_{k=1}^{r} \frac{B_{2k}}{(2k)!} \left(f^{(2k-1)}(n) - f^{(2k-1)}(0) \right) + Q_r \tag{74}$$

where[46]

$$Q_r = \int_0^1 \frac{B_{2r+1}(t)}{(2r+1)!} \sum_{i=0}^{n-1} f^{(2r+1)}(i+t)\,dt \tag{75}$$

which holds for any function f possessing the appropriate derivatives. (74) and (75) can be obtained from the integral form of the remainder terms for the ordinary Taylor series calculated for f and also for its derivatives using unit displacement from the series expansion point. (See [12] for details.) Applying this formula to the functions $f(x) = \log(b+x)$ and $f(x) = \log(a+b+x)$ and subtracting yields the formula

$$\begin{aligned} &\log(\mu_{0,n}(a,b)) = \\ &\quad \mathcal{F}[t \mapsto (t-\tfrac{1}{2})\log(t)] \\ &\quad + \sum_{k=1}^{r} \frac{B_{2k}}{2k(2k-1)} \left(-\frac{1}{(a+b+n)^{2k-1}} + \frac{1}{(b+n)^{2k-1}} + \frac{1}{(a+b)^{2k-1}} - \frac{1}{b^{2k-1}} \right) \\ &\quad + R_r \end{aligned} \tag{76}$$

with remainder

$$R_r = \int_0^1 \frac{B_{2r+1}(t)}{(2r+1)!} \left[-\psi^{(2r)}(t+a+b+n) + \psi^{(2r)}(t+b+n) + \psi^{(2r)}(t+a+b) - \psi^{(2r)}(t+b) \right] dt\,. \tag{77}$$

In this remainder term, $B_{2r+1}(t)$ is the Bernoulli polynomial [1, 804–6], and

$$\psi^{(2r)}(t) = \frac{d^{2r+1}}{dt^{2r+1}} \log(\Gamma(t)) \tag{78}$$

[43] either directly by subtraction, or by using (71) with $g(t) = (t+c)\log(t)$, $g''(t) = \frac{1}{t} - \frac{c}{t^2}$, monotonic decreasing in c

[44] This integer is the order of the $(1-P)$-moment. Does the E-M series which results (i.e. equations (76) and (77)) also hold exactly for non-integer n? For our purposes, this does not matter.

[45] Note that we do not use the E-M formula as given in [1, p806] since this contains errors.

[46] There are some alternative forms for the remainder in the Euler-Maclaurin summation formula. Note that, unlike some others which require $r \geq 1$, the form Q_r of remainder used here continues to be correct for $r = 0$ (provided that the 'empty sum' convention that $\sum_{k=1}^{0} \cdot = 0$ is used in (74)).

is the polygamma function [1, pp258–60]. The term in the integrand in square brackets is actually a shorthand, based on the basic polygamma recurrence relation

$$\psi^m(t+n) = \psi^m(t) + (-1)^m m! \left(\frac{1}{t^{m+1}} + \ldots + \frac{1}{(t+n-1)^{m+1}} \right) ,$$

for the expanded form (in which it was derived, as (75)).

The partial sum in (76) is a sum of alternating terms, since, for each term, the bracketed part $\mathcal{F}[t \mapsto t^{-(2k-1)}]$ is negative, and the even Bernoulli numbers (excluding B_0) are known to alternate in sign. We can show that, as one might hope to find, the sequence of *remainder terms* $\langle R_r \rangle$ of the approximation also alternates. To see this notice firstly that the square-bracketed term, $\xi(t)$, say, in (77) is a positive decreasing function of t. This fact is a consequence of putting $g = \psi^{(2r)}$ in (71) (and replacing b by $b+t$), since then we have $g'' = \psi^{(2r+2)}$ and the even polygamma functions[47] are known to be *increasing* (and negative) on the positive real axis. Alternatively, we obtain the same conclusion by leaving $\xi(t)$ in its original expanded form and rearranging the terms to give

$$\xi(t) = (2r)! \sum_{i=0}^{n-1} \left(\frac{1}{(t+b+i)^{2r+1}} - \frac{1}{(t+a+b+i)^{2r+1}} \right) ,$$

which is positive decreasing in t by the convexity of the inverse power function. The second requirement to deduce that $\langle R_r \rangle$ alternates in sign is the well-known property of the odd Bernoulli polynomials. [1, pp804–5] tells us that $B_{2r+1}(t)$ has a zero at $t = \frac{1}{2}$, has sign $(-1)^{r+1}$ on the interval $0 < t < \frac{1}{2}$ and satisfies the identity $B_{2r+1}(1-t) = -B_{2r+1}(t)$. Putting the above facts together we can deduce that the integrand in (77) has sign $(-1)^{r+1}$ on the interval $0 < t < \frac{1}{2}$, sign $(-1)^r$ on $\frac{1}{2} < t < 1$ and conclude that

$$\begin{aligned}
(-1)^{r+1} R_r &= \int_0^1 \frac{(-1)^{r+1} B_{2r+1}(t)}{(2r+1)!} \xi(t)\, dt \\
&> \int_0^{\frac{1}{2}} \frac{(-1)^{r+1} B_{2r+1}(t)}{(2r+1)!} \xi(\tfrac{1}{2})\, dt + \int_{\frac{1}{2}}^1 \frac{(-1)^{r+1} B_{2r+1}(t)}{(2r+1)!} \xi(\tfrac{1}{2})\, dt \\
&= \xi(\tfrac{1}{2}) \int_0^1 \frac{(-1)^{r+1} B_{2r+1}(t)}{(2r+1)!}\, dt \\
&= 0
\end{aligned}$$

Thus R_r has sign $(-1)^{r+1}$ and we can conclude that we have obtained bounds

$$\begin{aligned}
&\mathcal{F}[t \mapsto (t-\tfrac{1}{2}) \log(t)] \\
&\quad + \sum_{k=1}^{2s+1} \frac{B_{2k}}{2k(2k-1)} \left(-\frac{1}{(a+b+n)^{2k-1}} + \frac{1}{(b+n)^{2k-1}} + \frac{1}{(a+b)^{2k-1}} - \frac{1}{b^{2k-1}} \right) \\
&< \log\left(\mu_{0,n}(a,b)\right) \\
&\qquad < \mathcal{F}[t \mapsto (t-\tfrac{1}{2}) \log(t)] \\
&\qquad + \sum_{k=1}^{2s} \frac{B_{2k}}{2k(2k-1)} \left(-\frac{1}{(a+b+n)^{2k-1}} + \frac{1}{(b+n)^{2k-1}} + \frac{1}{(a+b)^{2k-1}} - \frac{1}{b^{2k-1}} \right) , \\
&\qquad\qquad \text{for } s = 0, 1, 2, \ldots . \qquad (79)
\end{aligned}$$

Note that the $s = 0$ case tells us that, in fact, $\mathcal{F}[t \mapsto (t-\frac{1}{2}) \log(t)]$ is a strict *upper bound* for $\log\left(\mu_{0,n}(a,b)\right)$. This enables a simple improvement to the right-hand side of (72).

Since we have now shown the sequence $R_0, R_1, R_2 \ldots$ to be alternating in sign, we have immediately

$$|R_r| \quad < \quad |R_r - R_{r+1}|$$

[47] meaning the functions (78) with $r \geq 1$

$$= \left| \frac{B_{2r+2}}{(2r+2)(2r+1)} \left(-\frac{1}{(a+b+n)^{2r+1}} + \frac{1}{(b+n)^{2r+-1}} + \frac{1}{(a+b)^{2r+1}} - \frac{1}{b^{2r+1}} \right) \right|$$

$$= \frac{(-1)^r B_{2r+2}}{(2r+2)(2r+1)} \left(\frac{1}{(a+b+n)^{2r+1}} - \frac{1}{(b+n)^{2r+-1}} - \frac{1}{(a+b)^{2r+1}} + \frac{1}{b^{2r+1}} \right)$$

$$< \frac{(-1)^r B_{2r+2}}{(2r+2)(2r+1)} \left(\frac{1}{b^{2r+1}} - \frac{1}{(a+b)^{2r+1}} \right) < \frac{(-1)^r B_{2r+2}}{(2r+2)(2r+1)b^{2r+1}} \tag{80}$$

as a crude bound on the error of our approximation, so that, summarising our findings about the remainder term, we can say that

$$0 < (-1)^{r+1} R_r < \frac{(-1)^r B_{2r+2}}{(2r+2)(2r+1)b^{2r+1}} \,. \tag{81}$$

In the numerical results presented in §6.2, we chose to work with $r = 3$ and $r = 4$ to give us our lower and upper bounds (respectively) on $\mu_{0,n}(a,b)$. With these numbers of terms in the series, (81) becomes

$$-\frac{1}{1188b^9} < R_4 < 0 < R_3 < \frac{1}{1680b^7} \tag{82}$$

Two further problems remained to be addressed in order to implement an algorithm. The first problem is that of the size of the error bounds when b is small. Although for high reliability software families we would probably not expect an asymptote in the distribution of $P\,|\,(a,b)$ at $P = 1$, we might nevertheless in our prior distribution for $\langle a,b \rangle$ wish to assign a very small quantity of probability to such values. For this reason we prefer to use a numerical algorithm for $\mu_{0,n}(a,b)$ which is able to cope well with values of b close to or even less than 1. Fortunately, there is a relatively painless solution to this requirement. Examination of the remainder term (77) leads one to conclude that for small b the remainder is largely accounted for by the Euler-Maclaurin series' comparative inability to approximate the first few terms of the original series $\sum_{i=0}^{n-1}(\log(b+i) - \log(a+b+i))$. This suggests removing these few terms from the sum, say remove the first j terms.

$$\log(\mu_{0,n}(a,b)) = \left(\sum_{i=0}^{j-1} \log(b+i) - \log(a+b+i) \right) + \log(\mu_{0,n-j}(a,b+j))$$

Then these removed terms can be calculated directly, and the Euler-Maclaurin approximation used only for the later part of the sum which is equal to $\log(\mu_{0,n-j}(a,b+j))$. For j large enough so that we have $b+j$ greater than about 5 or 6, the new Euler-Maclaurin remainder will be very small. To be precise, for $b+j > 5$ as we in fact used in §6.2, we have from (82) an Euler-Maclaurin remainder satisfying

$$-4.310{\times}10^{-10} < R_4 < 0 < R_3 < 7.619{\times}10^{-9} \,.$$

Using this approach we avoided ever having to use the expansion (76) with any value of $b \leq 5$.

The second of the two problems mentioned is purely computational and has to do with the avoidance of a loss of precision on subtraction of very similar numbers, which could occur in several places due to the multiple occurrences of differences of the form $\mathcal{F}$, and to the fact that a, b and n may differ by quite large orders of magnitude. Our approach to avoiding such problems for the $\mathcal{F}[t \mapsto (t-\frac{1}{2})\log(t)]$ term in (76) is to write it as

$$\mathcal{F}[t \mapsto (t-\tfrac{1}{2})\log(t)] = -a\log\left(\frac{a+b+n}{a+b}\right) + \left(b-\frac{1}{2}\right)\log\left(\frac{a+b}{b}\right) - \left(b+n-\frac{1}{2}\right)\log\left(\frac{a+b+n}{b+n}\right)$$

except when $b \geq 100a$ in which case we note that the second and third of these three terms will begin to become very similar. (They are both asymptotic to a as $\frac{a}{b} \to 0$.) Therefore, under this condition we replace each by a Taylor series approximation with a subtracted to give

$$\mathcal{F}[t \mapsto (t-\tfrac{1}{2})\log(t)] = -a\log\left(\frac{a+b+n}{a+b}\right) + h\left(a,\frac{a}{b}\right) - h\left(a,\frac{a}{b+n}\right)$$

where

$$h(x,y) = -y\left(\left(\left(\left(\left(\frac{y}{6}-\frac{1}{5}\right)y+\frac{1}{4}\right)y-\frac{1}{3}\right)y+\frac{1}{2}\right)\left(x-\frac{y}{2}\right)+\frac{1}{2}\right) .$$

Note also that the terms in the Euler-Maclaurin sum itself are *also* of the $\mathcal{F}$ form and so could likewise give rise to imprecision via subtraction. We avoid this problem by removing a factor a, rewriting (also for computational efficiency reasons) $\frac{1}{a}\left(1/(a+b)^{2k-1} - 1/b^{2k-1}\right)$ as a function of a^2, and $b(a+b)$, and similarly rewriting $\frac{1}{a}\left(-1/(a+b+n)^{2k-1} + 1/(b+n)^{2k-1}\right)$ as a function of a^2, and $(b+n)(a+b+n)$, to give, for $r = 4$

$$\sum_{k=1}^{4} \frac{B_{2k}}{2k(2k-1)}\left(-\frac{1}{(a+b+n)^{2k-1}} + \frac{1}{(b+n)^{2k-1}} + \frac{1}{(a+b)^{2k-1}} - \frac{1}{b^{2k-1}}\right)$$
$$= \quad a\left[l\big((b+n)(a+b+n)\,,\,a^2\big) - l\big(b(a+b)\,,\,a^2\big)\right] ,$$

where

$$l(x,y) = \frac{1}{12x} - \frac{3x+y}{360x^3} + \frac{5x\,(x+y)+y^2}{1260x^5} - \frac{7x\,(x+y)^2+y^3}{1680x^7} .$$

References

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions*. Dover, New York, 1970.

[2] 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.

[3] B. de Finetti. *Theory of Probability*, volume 2. Wiley, London, 1975.

[4] M. Henrion and B. Fischhoff. Assessing uncertainty in physical constants. *American Journal of Physics*, 54(9):791–8, 1986.

[5] D. Hunns and Wainwright. Software-based protection for Sizewell B: the regulator's perspective, September 1991.

[6] 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.

[7] B. Littlewood and L. Strigini. Validation of ultra-high dependability for software-based systems. *Comm. Assoc. Computing Machinery*, 36(11), November 1993.

[8] B. Littlewood and D. R. Wright. On a stopping rule for the operational testing of safety-critical software. In *Proc.* 25^{th} *Fault Tolerant Computing Symposium*, Pasadena, June 1995. IEEE.

[9] W. M. Miller, L. J. Morell, R. E. Noonan, S. K. Park, D. M. Nicol, and B. W. Murrill. Estimating the probability of failure when testing reveals no failures. *IEEE Transactions on Software Engineering*, 18(1), 1992.

[10] Requirements and Technical Concepts for Aeronautics. Software considerations in airborn systems and equipment certification. Technical Report DO-178B, Department of Defense, 1992.

[11] 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.

[12] M. Spivak. *Calculus*. World Student Series Edition. Addison-Wesley, Menlo Park, California, 1973.