Conservative Claims for the Probability of Perfection of a Software-based System Using Operational Experience of Previous Similar Systems

Xingyu Zhao, Bev Littlewood, Andrey Povyakalo, Lorenzo Strigini, David Wright

Centre for Software Reliability, City University London

{Xingyu.Zhao.1, B.Littlewood, A.A.Povyakalo, L.Strigini, D.R.Wright}@city.ac.uk

Abstract

We begin by briefly discussing the reasons why claims of *probability of non-perfection* (*pnp*) may sometimes be useful in reasoning about the reliability of software-based systems for safety-critical applications. We identify two ways in which this approach may make the system assessment problem easier. The first concerns the need to assess the chance of lifetime freedom from failure of a single system. The second concerns the need to assess the reliability of multi-channel software-diverse fault tolerant systems – in this paper, 1-out-of-2 systems. In earlier work (Littlewood and Rushby 2012, Littlewood and Povyakalo 2013) it was proposed that, in certain applications, claims for possible perfection of one of the channels in such a system may be feasible. It was shown that in such a case there is a particularly simple conservative expression for *system pfd* (probability of failure on demand), involving the *pfd* of one channel, and the *pnp* of the other. In this paper we address the problem of how to assess such a *pnp*. In previous work (Zhao 2015) we have addressed this problem when the evidence available is only extensive failure-free working of the system in question. Here we consider the case in which there is, in addition, evidence of the previous success of the software development procedures used to build the system: specifically, several previous similar systems built using the same process have exhibited failure-free working during extensive operational exposure.

KEY WORDS: *Fault-free software; probability of perfection; 1-out-of-2 system reliability; software diversity; operational experience; Bayesian modeling; limited prior belief; guaranteed-conservative reliability claims.*

Notation Table

Notation	Definition	First seen in
pfd_A	probability of failure on demand of channel A	(1.1)
pnp_B	probability of non-perfection of channel B	(1.1)
p_A	a given upper bound on pfd_A	(1.2)
α_A	the doubt in pfd_A smaller than the given bound p_A	(1.2)
p_B	a given upper bound on pnp_B	(1.2)
α_B	the doubt in pnp_B smaller than the given bound p_B	(1.2)
pfd_{sys}	probability of failure on demand of the whole 1oo2 system	(1.3)
A_i	a previous similar system i	(2.1)

P_i	*pfd* of the system A_i	(2.1)
n_i	the number of demands being executed by A_i	(2.1)
w_i	the number of demands causing failures by A_i	(2.1)
k	total number of similar products	(2.6)
$f(p\|\theta)$	a parametric family of distributions characterized by θ from which the *pfd* of each system A_i are randomly sampled.	(2.2)
θ	a vector of parameters characterize $f(p\|\theta)$	(2.2)
θ_{PP}	the probability mass at the origin of $f(p\|\theta)$, which is one of the parameters in the vector θ	(3.1)
$g(\theta_{PP})$	prior distribution of θ_{PP}	(3.1)
θ^*_{PP}	posterior probability of perfection seeing evidence	(3.1)
π	a given non-zero point on a two-point $f(p\|\theta)$ in which the other point is 0.	(3.2)
y	a given upper bound on θ_{PP}	(3.1.1)
α_θ	the confidence in θ_{PP} being smaller than the given bound y	(3.1.1)
α^*_θ	posterior confidence in θ_{PP} being smaller than y seeing evidence	(3.1.2)
y_1	a given upper bound on θ_{PP}	(3.2.1)
$\alpha_{\theta 1}$	the confidence in θ_{PP} being smaller than the given bound y_1	(3.2.3)
y_2	a given upper bound on θ_{PP}	(3.2.2)
$\alpha_{\theta 2}$	the confidence of θ_{PP} being in the range of $y_1 \leq \theta_{PP} < y_2$	(3.2.3)
$\alpha_{\theta 1+\theta 2}$	the confidence in θ_{PP} being smaller than bound y_2	(3.2.4)
$\alpha^*_{\theta 1}$	posterior confidence in θ_{PP} being smaller than y_1 seeing evidence	(3.2.3)
$\alpha^*_{\theta 1+\theta 2}$	posterior confidence in θ_{PP} being smaller than y_2 seeing evidence	(3.2.4)
R	the probability that a randomly selected product from the development process is not perfect and passed n tests.	(4.1.1)
$g_{<\theta_{PP},R>}$	the prior joint distribution of θ_{PP} and R	(4.1.3)
γ_θ	the confidence in θ_{PP} being smaller than the given bound y	(4.2.1)
γ_r	the confidence in R being smaller than the given bound r	(4.2.2)
M_i	the probability mass of the area i on the joint distribution $g_{<\theta_{PP},R>}$	Figure 8
γ^*_θ	the posterior confidence in θ_{PP} being smaller than bound y	(4.2.3)
$\gamma_{\theta 1}$	the confidence in θ_{PP} being smaller than the given bound y_1	(4.4.1)
$\gamma_{\theta 2}$	the confidence of θ_{PP} being in the range of $y_1 \leq \theta_{PP} < y_2$	(4.4.2)
r_U	a certain upper bound on R	(4.4.4)
$\gamma^*_{\theta 1}$	the posterior confidence in θ_{PP} being smaller than bound y_1	(4.4.5)

1. Introduction

In earlier papers we considered the problem of assessing the probability that a software-based system is "perfect" (Zhao 2015) – i.e. will not fail however long it operates – or is "quasi-perfect" (Zhao 2017) – i.e. is close enough to perfect for certain practical purposes. In that work we addressed this problem when the evidence available is only extensive failure-free working of the present system. Here we consider the problem of assessing probability of perfection when there is, in addition, evidence of the previous success of the software development procedures used to build the current system and similar earlier ones. Specifically, this evidence takes the form of extensive failure-free working of several previous similar products that were built using the same development procedures.

We begin by briefly providing the reader with a motivation for our interest in "probability of perfection". The material here repeats, in shortened form, the introductory discussions of (Zhao 2015, Zhao 2017).

1.1 Why is an assessment of "probability of perfection" useful?

Software-based systems are used in an increasing number of applications where their failures may be very costly, in terms of monetary loss or human suffering. As a result, such systems often have very high dependability requirements. Although it is not common for these requirements to be expressed (at least publicly) in a numerical form, a rare exception is the claimed 10^{-9} probability of failure on demand (*pfd*) for the combined control and instrumentation safety systems of the UK European Pressurised Reactor (UK EPR) (HSE 2013).

To *achieve* these kinds of ultra-high reliabilities is clearly a difficult task of design and implementation. The problems of *assessing* what has been achieved – so as to be sufficiently confident that a particular system is safe enough to use – seem to us to be even harder. Direct assessment via black-box operational testing, for example, would require infeasible times on test (Littlewood and Strigini 1993) (Butler and Finelli 1993) to support claims for ultra-high reliability (even in the unlikely event that there were no doubts about the representativeness of the operational profile used in the testing, and we could be certain that the test oracle was perfect).

In the research reported here, and similar earlier work, we consider a different approach to this difficult problem of assessment. The idea is that, instead of claims for *reliability* – failure rates, *pfd*s, etc – we make claims for *perfection.* This word comes, we realise, with extensive baggage: here it just means that a "perfect" system will not fail however much operational exposure it receives. If we assume that failures of a software system can occur if and only if it contains faults, it means that the system is "fault-free". Readers may think, reasonably, that we can never be certain that a system is perfect in this sense; but they may be prepared to accept that such perfection is possible. In the face of this uncertainty, we shall use "probability of perfection" as a measure of how good such a system is.

In fact, as has been has pointed out before (Bertolino and Strigini 1998, Rushby 2009, Littlewood and Rushby 2012), the traditional processes of software assurance, e.g. those performed in support of DO-178B ((RTCA 1992) the guidelines for the certification of safety-critical aircraft software), can be best understood as developing evidence of possible perfection, rather than ultra-high reliability. Indeed, claims for the perfection of some systems may be more intuitively plausible than claims for very high reliability, since the two claims would be supported by different types of evidence and reasoning. A claim for an extremely small failure rate seems to acknowledge that the system in question is unlikely to be perfect –

for example because of the complexity of its functionality – and resulting assessment of an extremely small number may not be believable. A claim for perfection, on the other hand, may be based upon evidence that the design is simple enough that the designers had a chance of "getting it right".

What benefits can we expect from this change of approach, to seek confidence that a software system is perfect, rather than reliable? There seem to be two ways in which this approach may make the system assessment problem easier. The first addresses the need to assess the chance of *lifetime* freedom from system failure; the second addresses the need to assess the reliability of multi-channel software-diverse fault tolerant systems, since such architectures are obvious candidates for these demanding safety-critical roles (and in some cases have been shown to have ultra-high reliability – after the fact – e.g. some systems in Airbus aircraft).

Consider first the problem of lifetime reliability of a critical on-demand[1] system, such as a safety shut-down system for a nuclear reactor or hazardous chemical process. Whilst requirements for such a system are often expressed in terms of probability of failure on (a single) demand (*pfd*), in fact for most systems what is really needed is a high confidence that at most a very small number of failures will occur over all demands in the expected life of the system. For some systems – e.g. nuclear reactor protection systems – this required number may be zero. A *pfd* claim is thus really in support of a *lifetime* claim.

The point here is that a probability of perfection directly addresses a lifetime claim: it is precisely the probability that the system will experience no failures, no matter how long its exposur*e* (number of demands over its life) (Bertolino and Strigini 1998, Strigini and Povyakalo 2013). Consider the following (artificial) example to illustrate this point. Let's say we have a system for which we expect 100 demands in its lifetime, and we want to be 99% confident that it will survive all these without failure. To obtain such confidence, we need the *pfd* to be no worse than about 10^{-4}. If we expected 1000 demands during its lifetime, we would need a *pfd* no worse than about 10^{-5} to be 99% confident of seeing no failures. For 10,000 demands, we need a *pfd* of about 10^{-6}, and so on. As the expected number of lifetime demands increases, the required *pfd* needed to be 99% confident of seeing no failures becomes more and more demanding – and thus so does the task of assuring that the requirement has been met.

In contrast, of course, we could be 99% confident of seeing no failures in *any* number of demands if we were 99% confident in perfection. If we could support such a possible perfection claim, the need for extremely extensive (possibly infeasibly so) testing to establish a very small *pfd* disappears.

The second reason we might wish to claim a probability of perfection arises from some recent work on design diversity, which has long been proposed as a promising way of achieving high dependability for software-based systems. The intuitive rationale is that if we force two or more systems to be built differently, their resulting failures may also be different. So if, in a 1-out-of-2 protection system (1-o-o-2 system), channel *A* fails on a particular demand, there may be a good chance that channel *B* will not fail. Thus, diversity in computer-based, safety-critical systems is popular in some industrial sectors (e.g. avionics, rail, nuclear), and mandated or highly recommended, for highly critical functions, by various standards and regulators (Wood and Belles 2010). Some of these systems have exhibited remarkable dependability in operation. For example, the safety-critical flight control systems of Airbus fleets have experienced massive operational exposure (Boeing 2015) with apparently no

[1] We shall use the terminology of on-demand systems for the rest of the paper, but much of what we say will also be applicable to continuously operating systems.

critical failure (note, however, that these continuously operating systems have a different architecture from the 1-o-o-2 *on-demand* systems we treat in this paper). Of course, an absence of accidents due to software failures could be due to extreme rarity of the latter (as these system are built to very stringent quality standards) rather than their having occurred and having been tolerated thanks to diversity. But experience gives no evidence *against* the current views that support the use of diversity (Littlewood, Popov et al. 2001).

This kind of evidence based on massive operational exposure is, of course, only available after the fact. Assessing the reliability of such a design-diverse system before it is deployed remains a very difficult problem. We know, from experimental work (Knight and Leveson 1986, Eckhardt, Caglayan et al. 1991) and theoretical modelling (Littlewood and Miller 1989) that we cannot claim with certainty that there is independence between the failures of multiple software-based channels of a system. Thus for a 1-o-o-2 on-demand system, if channel *A* fails on a randomly selected demand, this may increase the likelihood that the demand is a "difficult" one and so increase the likelihood that channel *B* also will fail. So even if we know the marginal probabilities of failures of the two channels, say pfd_A and pfd_B, from extensive testing, we cannot simply multiply them and claim the system *pfd* is $pfd_A \times pfd_B$.

In recent work by Littlewood and Rushby (Littlewood and Rushby 2012), the authors proposed a new way to reason about the reliability of a special kind of 1-o-o-2 on-demand system. Here channel *A* is conventionally engineered and presumed to contain faults, and thus supports only a *pfd* claim (say pfd_A). Channel *B* on the other hand is extremely simple and extensively analyzed, and thus is "possibly perfect"; in (Littlewood and Rushby 2012) the claim about this channel is a probability of non-perfection, pnp_B[2]. They show that:

$$\Pr(\text{system fails on randomly selected demand} \mid pfd_A, pnp_B) \leq pfd_A \times pnp_B \tag{1.1}$$

The result depends on the events "*A* fails on a randomly selected demand" and "*B* is not perfect" being conditionally independent, given that the probabilities of these events, respectively pfd_A and pnp_B, are known. The right hand side of equation (1.1) is a conservative bound for the system's probability of failure on demand (pfd_{sys}). The conservatism arises from an assumption that if *B* is imperfect, it always fails when *A* does.

The result is useful because it allows multiplication of two small numbers to obtain (a bound on) pfd_{sys} – hopefully a *very* small number. That is, it overcomes the problem with the unattainable result above, involving the product of the small numbers pfd_A and pfd_B, which requires the unreasonable assumption of independence of channel failures.

In reality, of course, an assessor would not know pfd_A and pnp_B with certainty: there is epistemic uncertainty about their numerical values. In principle, an assessor could represent his uncertainty here via a (bivariate) distribution for the two unknowns. In practice people find this kind of thing very difficult, if not impossible. Whilst assessors may be able to make informed statements about their marginal beliefs about the two parameters separately, they will usually be unable to say anything about their dependence.

In (Littlewood and Povyakalo 2013) this problem is addressed: results are obtained that require *only* an assessor's marginal beliefs about the individual numbers, i.e. they do not require the assessor to say anything about dependence between the two numbers. For

---

[2] In the present paper we shall usually deal with "probability of perfection", which is of course simply 1- *pnp*. Readers will find a discussion a little later on about precisely what "probability" means in this novel context, concerning perfection (we take it for granted that its meaning in the term "probability of failure on demand" is well-understood).

example, Theorem 5 of (Littlewood and Povyakalo 2013) shows that if an expert has a single (marginal) confidence bound for each channel parameter

$$\begin{aligned} \Pr(pfd_A < p_A) &= 1-\alpha_A \\ \Pr(pnp_B < p_B) &= 1-\alpha_B \end{aligned} \tag{1.2}$$

then the following is a conservative confidence bound for the system *pfd*:

$$\Pr(pfd_{sys} < p_A \times p_B) > 1-(\alpha_A + \alpha_B) \tag{1.3}$$

In words, the system claim $p_A \times p_B$ is the product of the channel claims, and the doubt about the truth of this system claim is the sum of the doubts, $\alpha_A + \alpha_B$, about the channel claims.

To summarise, the results of (Littlewood and Rushby 2012) and (Littlewood and Povyakalo 2013) reduce the problem of assessing the *pfd* of this kind of special 1-o-o-2 system to one concerning simply marginal beliefs about the parameters pfd_A and pnp_B. In particular, we do not need to be concerned about aleatory dependence between failures of *A* and *B* (Littlewood and Rushby 2012), nor about epistemic dependence between beliefs about model parameters (Littlewood and Povyakalo 2013). The price paid, of course, in each case, is conservatism in the results.

There is a large literature on the assessment of *pfd* from statistical analysis of operational tests – see e.g. (Littlewood and Wright 1997) – so the first of these parameters could be easily assessed, e.g. in terms of a Bayesian posterior distribution based on evidence from operational testing.

That leaves pnp_B, the assessment of which is the subject of the current paper.

1.2 How can we assess *pnp*?

We first digress briefly here to discuss what "probability" means in the phrase "probability of non-perfection". The aim here is to suggest how we might learn about its magnitude, and thus motivate the detailed modeling in the body of the paper.

Clearly, each program either is, or is not, "perfect". An observer of a particular program may be uncertain whether it is perfect or not, so one interpretation of this probability is the observer's subjective strength of belief in non-perfection. A useful classical (frequentist) interpretation can be obtained from the notion of software development as the random selection of a program from a population of programs, as first introduced in (Eckhardt and Lee 1985). The idea here is that, for a particular problem that a program is to solve, and a particular development process to be used in creating the program, there is a hypothetical population of all programs that could be written to solve the problem using the process, and a distribution over this population that determines the probability of selection for each member of the population (note that, almost certainly, such a distribution will *not* result in equi-probable selection of programs).

A slightly different conceptual formulation is to consider the population of all programs that could be written. A particular problem to be solved would then be completely characterized by its probability distribution over all programs. In this scenario, many programs (in fact almost all) would have zero probability of selection.

In these rather abstract models of software development *pnp* is a property of a hypothetical population of programs: each program in this population will be either perfect, or not, for the problem being solved. The act of randomly selecting a single program from the population – i.e. developing a program to solve the problem – results in a program that is either perfect or

not. The parameter *pnp* is just the probability that such a randomly selected program is not perfect.

Notice how this contrasts with the interpretation of *pfd*, which concerns a *single* program. Here there is a distribution over the population of all demands – the "operational profile". Each demand will be executed correctly, or not, and the act of selection results in a demand that is correctly, or incorrectly, executed. The parameter *pfd* is then just the probability that a randomly selected demand is one of those that cannot be executed correctly.

This difference in interpretation of the two probabilities hints at the different ways we might learn about them. So, for a *pfd* concerning a particular *program*, we would like to see the outcome (success or failure) of many randomly selected demands executed by that program. But for a *pnp* concerning a development *process* (population of programs), we would like to know, for each of many randomly selected programs, which of them are not perfect.

It is this observation that underpins much of the rest of this paper: the model we develop allows an assessor to learn about the quality of a development process by observing previous products developed to solve a similar problem, and thus to learn about the *pnp* for that process/problem pair.

There is, of course, an important difference between this learning about *pnp* and learning about *pfd*. For the latter, when we see many demands, we shall *know* for each demand whether it was a success or a failure. In contrast, for previous software products, we shall not know with certainty which were perfect; rather at best we may only have evidence of extensive failure-free operation.

Our early work on this problem, e.g. (Zhao 2015), used only evidence of perfect working on many statistically representative demands on the present single system. In practice, of course, other kinds of evidence will be available and should be used to increase confidence in claims for perfection. In the current paper we consider the case where, in addition to evidence of failure-free working on *this* system, there is also evidence of failure-free working of previous, similar systems. The idea here is that such evidence supports claims about the quality of the development procedures used to build this system and previous ones. This kind of process-quality evidence has long been used in support of claims about system dependability – "trust us to have built *this* system right because we have built similar ones in the past, and here's the evidence of their success." But hitherto this kind of claim has been supported only in informal, qualitative ways; we present here what we think is the first formal mathematical model to take account of such evidence in support of quantitative perfection claims.

Our approach here – as in previous work – will be Bayesian. As is well known, a serious problem in Bayesian analysis concerns the elicitation of prior beliefs from the expert who "owns" the problem. For the safety-critical systems that motivated our work, this expert assessor may be a regulator, or perhaps a safety engineer, employed by a licensee of regulated plant such as a nuclear power station.

Ideally, the Bayesian approach requires a complete description of the subjective prior uncertainty about the parameter(s) of a model. In the case of a model with a single parameter, this involves a complete uni-variate distribution representing the expert's subjective beliefs about the parameter. More generally – and posing greater difficulties – it requires the expert to provide a complete multi-variate distribution representing his beliefs about several parameters.

Experts find it difficult, or impossible, to provide such complete distributions for their prior beliefs, particularly for the cases we are examining, which concern difficult concepts like possibility of perfection of a complex system. Any practical approach to these problems

needs to take account of these difficulties, and this is a major theme of the work reported here and in our previous work (Bishop, Bloomfield et al. 2011, Zhao 2015). In this paper we shall assume that the expert can provide only quite limited beliefs, e.g. only one or two percentiles of a distribution rather than the complete distribution. We show how to do this in such a way that useful results are still obtained. The modeling trade-off here is between the restricted prior beliefs being too minimal to be useful, and their being so extensive as to place impossible demands upon the expert.

Throughout our modeling here, as in previous work, the approach is one that guarantees that the eventual top-level claims about a system's dependability are *conservative*, as seems necessary for the safety-critical applications that we have in mind. We do this as follows. For any particular restricted set of expert prior beliefs – e.g. one or two percentiles in the case of prior beliefs about a single parameter – there will be an infinite number of complete prior distributions satisfying these constraints. We can think of any one of these as satisfying the expert's limited expressed beliefs. In section 4.2 below, we extend essentially the same idea to the case of *two* unknown parameters, and we believe that, in principle, the same underlying reasoning may be found applicable to any vector of unknown system parameters. For two parameters, this is not necessarily equivalent merely to eliciting one or more marginal percentiles for each parameter in turn, since the expert might believe in an association between them (a form of subjective belief that has elsewhere been termed "epistemic correlation") which he might also wish to express, in some limited way. In either case, whether for a single parameter or for multiple unknown parameters, we show how to choose the single constraint-complying prior (whether univariate or multivariate) that provides the most conservative posterior beliefs about the probability of (im)perfection (Strictly speaking, there may be more than one prior that produce this same result -but there is none that is more conservative than the one that we present in a later section). Use of this extreme prior then guarantees conservatism of the final results. These priors turn out to be ones that have non-zero probability mass at only a few points – see also (Bishop, Bloomfield et al. 2011).

2. A general doubly stochastic hierarchical model of "process quality" evidence

We begin by describing a general model for learning about the dependability of a system using evidence about the efficacy of the development procedures used to build it, obtained from observing failure-free operation of previous "similar" systems, built "similarly". We shall then later specialize this general model in order to support claims about perfection.

The informal scenario we have in mind is the following. We have built a new system. We want to predict its reliability (including its possibility of perfection). We have some information about this system, e.g. it has survived a number of tests without failure. We also have similar evidence from previous similar systems. We wish to use all this evidence to predict the reliability of our current system. The informal idea here is that there is a common development process for all systems that are addressing "similar" problems. The process thus can be thought of as generating a possible *population* of systems, of which the current one – the one that interests us – is an instance. Knowledge of the population thus informs our judgment of *this* system.

More particularly, and more formally, we shall consider here the case where the "evidence" for each system is just the record of outcomes (success or failure) for the demands that it has experienced during its lifetime: see Figure 1. This evidence thus supports the kind of informal claim (see Section 1) that we have seen used by builders of safety-critical systems: "trust us to have built *this* system right because we have built similar ones in the past, using similar

procedures, and here's the evidence of *their* success, i.e. evidence of the success of our design-and-build process that was used to build *this* system."

Our aim is to formalize such reasoning, and to do so we shall begin by assuming the following doubly stochastic model.

At the first level of the model, for each system, A_i, the successes/failures on successive demands will form a Bernoulli process. That is, successive demands of system A_i fail independently, with probability P_i, say. Thus, if system A_i has executed a known number n_i of demands, and $P_i=p_i$, the (random variable) number of failures W_i has the binomial distribution Bnl(p_i, n_i) given $P_i=p_i$:

$$W_i \mid n_i, p_i \sim \binom{n_i}{w_i} p_i^{w_i} (1-p_i)^{n_i - w_i} \tag{2.1}$$

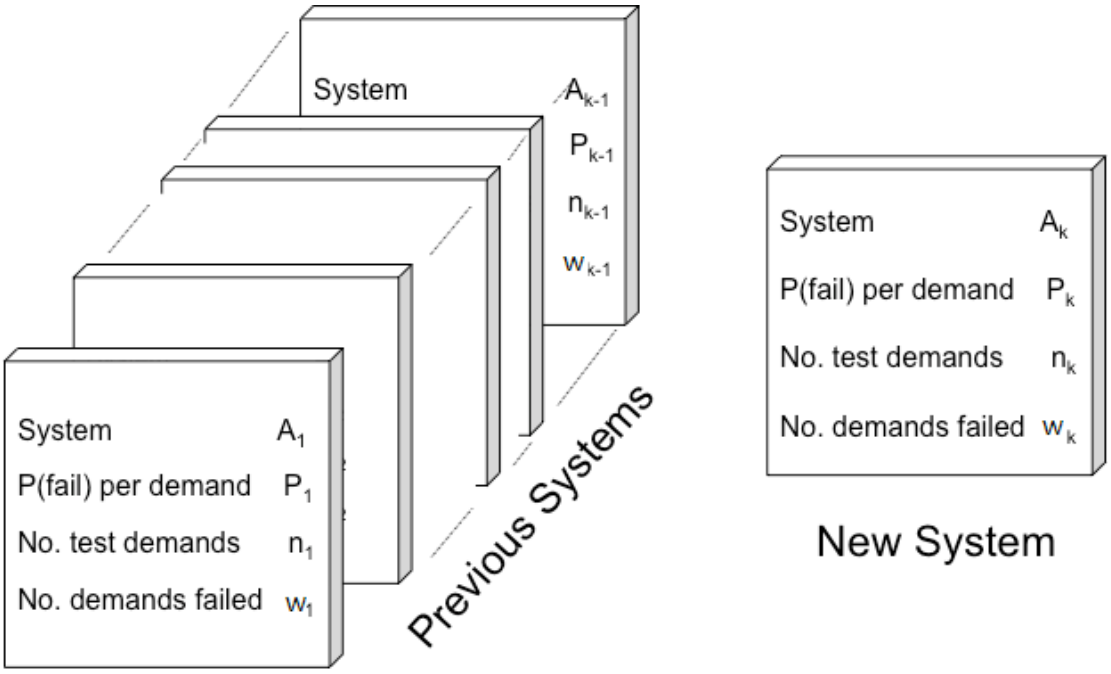

Figure 1. Evidence from (k-1) previous "similar" systems' behaviour to support a claim about the k-th novel (but "similar") system.

The process generates different systems, A_i, and these different systems will have different *pfd*s, p_i. In this second level of the doubly stochastic model, we assume that these system failure probabilities are sampled independently from a distribution, f. More precisely, the P_i are independent, identically distributed random variables

$$P_i \mid \theta \sim f(p \mid \theta) \tag{2.2}$$

from some parametric family of distributions $f(p \mid \theta)$ that is characterized by the parameter θ (which may be a vector).

A distribution f from this family (i.e. having a particular, but likely unknown, value of the parameter θ) can be thought of as characterizing the development process, as far as its ability to produce reliable systems is concerned. More precisely, it characterizes the reliability variation between systems that come from the same development process (i.e., that have the

same value of the parameter θ). Thus if it were a highly concentrated distribution (i.e. has a very small variance), the process could be regarded as *consistent*, inasmuch its products would have reliabilities that differed little from one another. If most of the probability mass of the distribution *f* were near the origin, the system probabilities of failure would all tend to be small – i.e. the process would be "a good one", producing mainly reliable products.

In this reliability model, building a system using this particular process can be thought of as randomly selecting a *pfd* from a population with distribution *f*. We shall be interested here in the case where there is a possibility of a chosen system being perfect – i.e. its *pfd* is zero. In fact our interest will centre on the probability of this event, which is just the probability mass of *f* concentrated at the origin.

Initial uncertainty about inter-product variation in this Bayesian model is represented by the prior distribution for the parameter θ:

$$\Theta \sim \text{Prior}_\theta(\theta) \tag{2.3}$$

Figure 2 shows the two-stage (what we have called doubly stochastic) dependency model. The only random variables (those with upper case letters in the figure) that are observable here are the W_i. Consider, however, the following thought experiment. First, for each particular system A_i let $n_i \to \infty$; it is easy to see that the ratio w_i/n_i converges to the true (but unknown, unobservable) probability of failure on demand, p_i. Now imagine generating k systems – i.e. randomly and independently selecting k probabilities of failure on demand, $\{p_i\}$, from f – and let $k \to \infty$. This generates an infinite number of *systems*, and thus an infinite random sample of *pfd*s, p_i. A histogram of these p_is will converge to the true (but unknown) distribution $f = f(p|\theta)$, and we shall therefore know the value of θ with certainty.

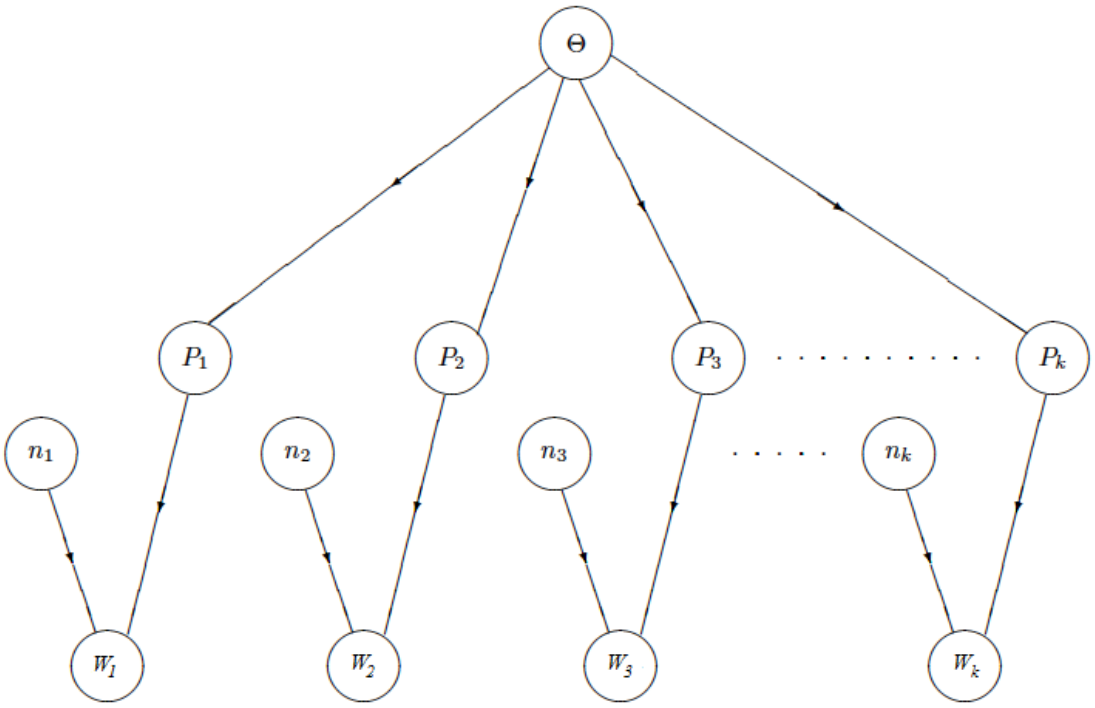

Figure 2. The dependency model. Note that the realisations of the random variables $\{W_i\}$(Zhao 2015, Zhao 2017) will be observable here; those of $\{P_i\}$ and Θ will not.

We have defined here a model in which the parameter θ can be thought of as characterizing a family of systems generated by a particular process. For a system chosen at random from a particular family (i.e. particular θ) and observed for n demands, it follows that (W, P) has the joint distribution

$$(W,P)\,|\,n,\theta \sim \binom{n}{w} p^{w}(1-p)^{n-w}\, f\,(p\,|\,\theta) \tag{2.4}$$

given n and θ. Integrating (2.4) over p gives

$$W\,|\,n,\theta \sim \binom{n}{w} \int_0^1 p^{w}(1-p)^{n-w}\, f\,(p\,|\,\theta)dp \tag{2.5}$$

Informally, our uncertainty about the process is just uncertainty about f. This in turn is just uncertainty about θ. So we need to "learn" about θ, which we can do via Bayes' Theorem when we have collected evidence from the testing of multiple systems. So, if for k systems we have seen w_i failures in n_i demands (i=1, 2, ..., k), the likelihood function for θ is

$$L(\theta;\langle n_i, w_i\rangle_{i=1}^{k}) = \prod_{i=1}^{k}\binom{n_i}{w_i}\int_0^1 p^{w_i}(1-p)^{n_i-w_i}\, f\,(p\,|\,\theta)dp \tag{2.6}$$

from which, with (2.3), we can obtain the posterior distribution for θ.

Rather than continue this account with the full general model, we shall now illustrate the ideas via a very simple example. In particular, here the parameter θ will be specialised to represent the *probability of perfection*, upon which our interest centres.

3. A simple example using two-point distribution as $f(p|\theta)$

We consider a simple case of a two-point distribution of $f\,(p|\theta)$ with θ_{PP} mass at origin (representing probability of perfection) and $1-\theta_{PP}$ mass at π, as in Figure 3.

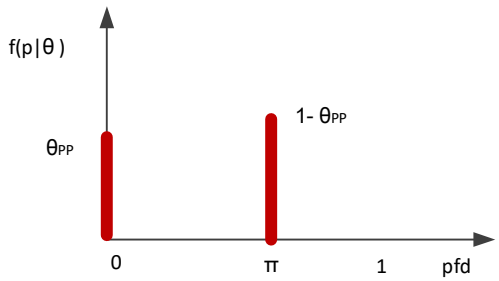

Figure 3 two-point *f* distribution

Here, for simplicity, π is assumed to be known. This is, of course, unrealistic, but will suffice for this simple illustration. It follows that the parameter θ_{PP}, representing probability of perfection, on its own completely characterises the distribution f. That is, θ_{PP} alone is the parameter θ in the previous section, at the process level, that determines the development process quality.

Our interest is in the *posterior* distribution of probability of perfection, say θ_{PP}^{*}, having seen evidence from the successful operation of several products as shown in Figure 1. If an expert were able to provide a complete prior distribution, say $g(\theta_{PP})$, he can simply use Bayes' theorem to obtain this posterior distribution. Assuming the k systems have each executed n

demands without failure (we refer to this as the "process evidence" in what follows, for notational simplicity), we have

$$\theta_{PP}^{*} \sim g(\theta_{PP}|\text{process evidence}) = \frac{L(\theta_{PP};\text{process evidence})g(\theta_{PP})}{\int_0^1 L(\theta_{PP};\text{process evidence})g(\theta_{PP})d\theta_{PP}} \quad (3.1)$$

where the likelihood function is:

$$L(\theta_{PP};\text{process evidence}) = [\theta_{PP} + (1-\pi)^n(1-\theta_{PP})]^k \quad (3.2)$$

As we have argued elsewhere (Bishop, Bloomfield et al. 2011, Zhao 2015) it is generally unreasonable to expect an expert to be able to provide a complete $g(\theta_{PP})$ as prior. But it is often feasible to express some *precise* but *partial* beliefs about the unknown θ_{PP}. We now examine, as an example, the very simple case of knowing only one percentile of θ_{PP}.

3.1 One percentile of θ_{PP} as prior partial belief

In this case, the assessor can only tell us one percentile of θ_{PP} as his prior belief, i.e. the confidence bound (y, α_θ):

$$P(\theta_{PP} < y) = \alpha_\theta \quad (3.1.1)$$

To use the results of (Littlewood and Rushby 2012, Littlewood and Povyakalo 2013) – see equations (1.2) and (1.3) – we need a *posterior* confidence bound. If we had a complete prior distribution, this would be

$$P(\theta_{PP} < y \mid \text{process evidence}) = \alpha_\theta^{*} = \frac{\int_0^{y^-} L(\theta_{PP};\text{process evidence})g(\theta_{PP})d\theta_{PP}}{\int_0^1 L(\theta_{PP};\text{process evidence})g(\theta_{PP})d\theta_{PP}} \quad (3.1.2)$$

where g is the prior distribution as before, and L is the likelihood function in equation (3.2).

Now in general there will be an infinite number of prior distributions satisfying (3.1.1). The question is which of these gives us the most conservative (i.e. maximum[3]) $\alpha_\theta{}^{*}$.

We can show that (see Appendix 1 for proof) all the mass of the most conservative prior $g(\theta_{PP})$ collapses to the point y, as in Figure 4. Therefore, starting with only one percentile prior (3.1.1), we cannot learn about θ_{PP} from process evidence, i.e. $\alpha_\theta = \alpha_\theta{}^{*}$.

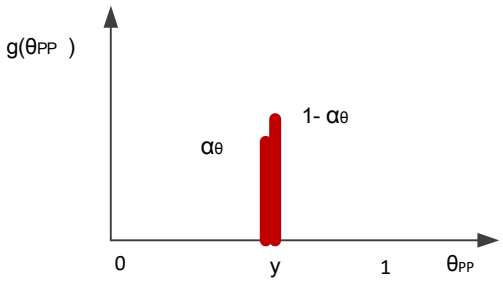

Figure 4 the most conservative prior of $\boldsymbol{\theta_{PP}}$

This case is, of course, unhelpful. The extremely minimal prior belief here is *too* minimal to provide useful learning from the multiple product evidence. We therefore now consider the case in which the expert can provide *two* percentiles of θ_{PP} as prior belief.

---

[3] $P(\theta_{PP}<y)=\alpha_\theta$ equals to $P(1-\theta_{PP}\leq 1-y)=1-\alpha_\theta$, so α_θ is essentially the doubt that *pnp* (probability of non-perfection) is smaller than a certain bound. We want this doubt to be small, so it is conservative to maximize it.

3.2 Two percentiles of θ_{PP} as prior belief

Consider the case where we have two percentiles of $g(\theta_{PP})$, i.e. the confidence bounds $P(\theta_{PP} < y_1) = \alpha_{\theta 1}$ and $P(\theta_{PP} < y_2) = \alpha_{\theta 1+\theta 2} = \alpha_{\theta 1} + \alpha_{\theta 2}$, as in Figure 5. The corresponding posterior confidence bounds in terms of a complete prior distribution g are:

$$P(\theta_{PP} < y_1 \mid \text{process evidence}) = \frac{\int_0^{y_1^-} L(\theta_{PP};\ \text{process evidence})\, g(\theta_{PP})\, d\theta_{PP}}{\int_0^1 L(\theta_{PP};\ \text{process evidence})\, g(\theta_{PP})\, d\theta_{PP}} \tag{3.2.1}$$

$$P(\theta_{PP} < y_2 \mid \text{process evidence}) = \frac{\int_0^{y_2^-} L(\theta_{PP};\ \text{process evidence})\, g(\theta_{PP})\, d\theta_{PP}}{\int_0^1 L(\theta_{PP};\ \text{process evidence})\, g(\theta_{PP})\, d\theta_{PP}} \tag{3.2.2}$$

where again L is the likelihood function (3.2).

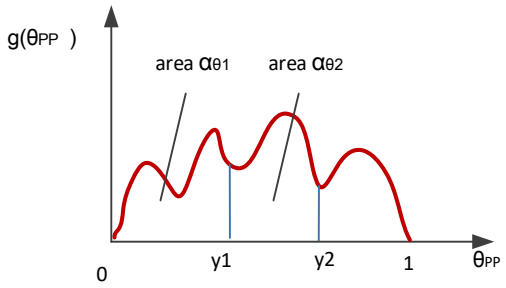

Figure 5 Two percentile constraints of θ_{PP} as priori belief

We can show (see Appendix 2) that the most conservative prior distributions satisfying the two expert belief constraints are the point-mass distributions in Figure 6. The corresponding most conservative posterior confidence results, $\alpha^*_{\theta 1}$ and $\alpha^*_{\theta 1+\theta 2}$ (keeping the same bounds, y_1 and y_2) are:

$$\alpha^*_{\theta 1} = \frac{\alpha_{\theta 1}}{\alpha_{\theta 1} + \alpha_{\theta 2} + \dfrac{L(y_2)(1-\alpha_{\theta 1}-\alpha_{\theta 2})}{L(y_1)}} \tag{3.2.3}$$

$$\alpha^*_{\theta 1+\theta 2} = \frac{L(y_1)\alpha_{\theta 1} + L(y_2)\alpha_{\theta 2}}{L(y_1)\alpha_{\theta 1} + L(y_2)(1-\alpha_{\theta 1})} \tag{3.2.4}$$

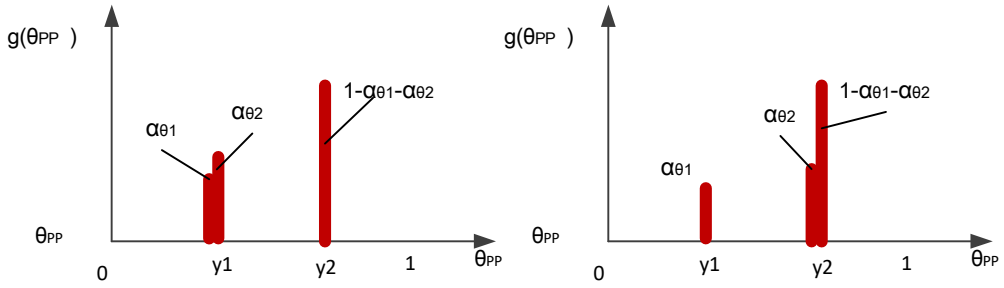

Figure 6 The most conservative prior distributions giving results (3.2.3) and (3.2.4) respectively

Table I gives some numerical results based on different values of the various model parameters.

Consider case #1, with some arbitrary but not unreasonable numbers for the parameters. Here we do learn something from the process evidence (10 similar products that each passed 10000 tests). However, even though the results are better than the "learn-nothing" results in section 3.1, the two posterior confidence bounds on θ_{PP} are still very unhelpful: the decrease of doubts from priors to posteriors (i.e. comparing the columns $\alpha_{\theta 1}$ with $\alpha^*_{\theta 1}$ and $\alpha_{\theta 1+\theta 2}$ with $\alpha^*_{\theta 1+\theta 2}$) is quite limited.

Table I Numerical examples of the case with prior beliefs expressed in terms of two percentiles

Case #	y_1	$\alpha_{\theta 1}$	y_2	$\alpha_{\theta 1+\theta 2}$	π	n	k	$\alpha^*_{\theta 1}$	$\alpha^*_{\theta 1+\theta 2}$
1	0.9	0.05	0.99	0.1	0.001	10000	10	0.020540109	0.071473861
2	0.9	0.05	0.99	0.1	0.001	10000	50	0.000472913	0.053056233
3	0.9	0.05	0.99	0.1	0.001	1000000	10	0.02053921	0.071473018
4	0.9	0.05	0.99	0.1	0.01	10000	10	0.02053921	0.071473018
5	0.5	0.05	0.7	0.1	0.001	10000	10	0.001913788	0.054352684
6	0.9	0.05	0.99	0.1	0.000001	10000	10	0.04959824	0.099598419
7	0.9	0.05	0.99	0.1	0.000001	10000	50	0.048019803	0.098024151
8	0.9	0.05	0.99	0.1	0.000001	1000000	50	0.002897123	0.055239721
9	0.9	0.05	0.99	0.1	0.000001	1000000	10	0.029020956	0.079498839

Cases #2 and #3 consider the impact of increases in *k* and *n*, respectively. The benefit of increasing *k* (from 10 to 50) is much greater – two orders of magnitude – than increasing *n* (from 10^4 to 10^6), even though the total number of tests (i.e. the number *k***n*) for case #2 is 500000 which is much less than the 10000000 tests in case #3. The latter observation should not be taken as a guide to the relative "cost" of the two scenarios, however, since the number of test cases is a poor guide to the "cost" of gaining the evidence in these two cases: developing many systems is far more onerous than generating many test cases.

The importance of *k* here is not surprising. It seems intuitively obvious that "all things being equal" we learn more about the efficacy of the process by having evidence of good working from many products, than we do from massive exposure of only a few products. In practice, of course, it seems unlikely there will be available very many previous products to provide this kind of evidence.

Another observation that confirms intuition is that the weaker the claims are, the greater the confidence we could obtain from process evidence. Consider case #5: the claim $\theta_{PP} < 0.5$ is weaker than it is for the previous cases ($\theta_{PP} < 0.9$), so seeing the same evidence (*n*=10000,

k=10), we learn more in this case. Notice also that for each case the improvement for $\alpha^*_{\theta 1}$ is better than for $\alpha^*_{\theta 1+\theta 2}$. Again, this is because the claims (i.e. $P(\theta_{PP} < y_1)$) for $\alpha^*_{\theta 1}$ are weaker than the ones for $\alpha^*_{\theta 1+\theta 2}$ (i.e. $P(\theta_{PP} < y_2)$).

We shall not pursue analysis of Table I any further here, since this set-up with the two-point *f* distribution is unlikely to be a realistic representation of reality; it was meant only to be an illustrative aid to understanding.

However, the reader will note that, even for this unrealistically simple example, there are differences between the results of Sections 3.1 and 3.2. The *very* restricted prior beliefs of Section 3.1 do not allow any useful learning from the evidence of failure-free working of multiple systems; the slightly more informative prior beliefs of Section 3.2, in contrast, provide some modest learning. An important issue in this kind of study is identifying exactly *how* minimal prior beliefs can be – to aid the task of the assessor – without being *too* minimal to give useful results. In the following section we introduce a more plausible model to investigate such issues.

4. More practical assumption of an arbitrary *f* distribution with mass at the origin

In Figure 7 we show a schema of the different layers of our model that need to be populated to enable its use.

At the bottom is the "behaviour level" that describes the *aleatory uncertainty* about failures or successes on demand of a system. Each system has a *pfd* that determines whether a randomly selected demand will result in failure (or not), and successive demands are assumed to fail independently – all as described in Section 2.

This *pfd* of a (randomly chosen) *system* is determined at the middle "product level" of Figure 7: for different products, these *pfd*s are independent and identically distributed random variables from a distribution $f(p|\theta)$, characterized by a (possibly vector) parameter, θ.

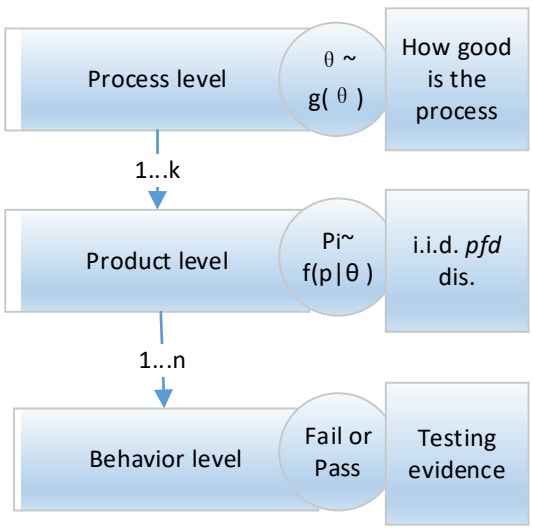

Figure 7 Schema for the different stages of our model.

This parameter, θ, is unknown. Our statistical inference – based upon evidence from the testing of several products – concerns this parameter at what we have called the "process level" of the Figure 7 schema. This statistical inference will proceed automatically, via Bayes' Theorem, starting from the expert's prior distribution, $g(\theta)$.

In the previous section, to illustrate the general approach, we showed an analysis based on an unrealistic two-point-mass distribution *f* at the middle level of the schema; in this section we propose a way to treat this middle layer of the model more realistically. Readers will note, however, that even with the unrealistic "product level" simplification of Section 3 – where the only unknown quantity is θ_{pp} – there arise quite difficult problems concerning prior beliefs.

What would be a reasonable assumption about *f*? Clearly, this must have mass at the origin, θ_{pp}, since this is the centre of our interest. But what happens in (0,1]? Having probability mass only at the point π – with no probability density anywhere else in the interval – as in Section 3, is clearly unbelievable. In fact non-zero probability mass at any point in (0,1] seems unbelievable. Most experts, we think, would be prepared to say that *f* had an absolutely continuous density for positive values of *pfd*.

An obvious "classical" approach then would be to assume a parametric family for this density. E.g. we could assume that *p* had a (conditional, given $p>0$) Beta distribution with parameters (α,β). In that case the unknown model parameter at the top level of the Figure 7 schema would be $\theta=(\theta_{pp},\alpha,\beta)$.

There are problems with such an approach, however. Most importantly, it is hard to justify the choice of a Beta (or any other) parametric family. Additionally, experts would find it hard to express their prior beliefs about the vector $(\theta_{pp},\alpha,\beta)$. In what follows, therefore, we propose a different way forward that does not rely on such a parametric assumption; in fact no assumptions are made about the shape of the distribution *f* for non-zero *p*.

4.1 The introduction of *R*

We consider an unknown arbitrary distribution *f* with some mass (θ_{PP}) at the origin. Our objective function is:

$$P(\theta_{PP} < y|\ \text{process evidence}) = \frac{P(\theta_{PP} < y \text{ and process evidence})}{P(\text{process evidence})}$$
$$= \frac{E_f(I_{\theta_{PP}<y}(f) \times P(\text{process evidence}|f))}{E_f(P(\text{process evidence}|f))}$$

where $I_{\theta_{PP}<y}(f)$ is an indicator function using all possible *f* as inputs. When the mass at the origin of a *f* is less than *y*, $I_{\theta_{PP}<y}(f) = 1$, otherwise 0. E_f is the expectation over all possible *f*. The use of the indicator function $I_{\theta_{PP}<y}(f)$ ensures the mean value in the numerator only involves those possible *f* having a mass at the origin less than *y*.

Now[4]

$$P(\text{process evidence}|f) = \left[\theta_{PP} + \int_{0+}^{1} (1-p)^n f(p|\theta)\, dp\right]^k$$

[4] The notation of 0+ in the following integral means the integral over the set which is the half open interval (0,1]; and similarly for y- in the later sections of the paper

So if we denote[5]

$$R = \int_{0+}^{1} (1-p)^n f(p|\theta)\, dp \tag{4.1.1}$$

then

$$P(\text{process evidence}|f) = [\theta_{PP} + R]^k \tag{4.1.2}$$

which is a function only of the pair $< \theta_{PP}, R >$. And in principle the joint distribution of θ_{PP} and R, say $g_{<\theta_{PP},R>}$, could be calculated from the $g(\theta)$ (i.e. the distribution of all possible *f*).

$$P(\theta_{PP} < y|\ \text{process evidence}) = \frac{E_f\left(I_{\theta_{PP}<y}(f) \times P(\text{process evidence}|f)\right)}{E_f\left(P(\text{process evidence}|f)\right)}$$

$$= \frac{E_{<\theta_{PP},R>}\left(I_{\theta_{PP}<y}(\theta_{PP}) \times [\theta_{PP} + R]^k\right)}{E_{<\theta_{PP},R>}([\theta_{PP} + R]^k)} \tag{4.1.3}$$

which depends only on the joint distribution of θ_{PP} and R, say $g_{<\theta_{PP},R>}$.

The importance of this result is obvious. The problem of how to deal with *f* at the "product level" of Figure 7 has been transformed into a problem at the "process level". Most importantly, no assumptions about the shape of *f* for non-zero *p* have been made. Instead an expert "only" has to express joint prior beliefs about the pair of parameters θ_{PP} and *R*. Mathematically, *R* is sufficient to represent the non-zero part of *f* in terms of Bayesian learning about the probability of perfection (Note the similarity of this observation to the concept of *sufficient statistic* in classical statistics; see, for example (Lehmann and Cassella 2003)).

We use quotes for "only" in the previous paragraph because this remaining problem concerning prior beliefs is not a simple one. Certainly, it would be unreasonable to expect an expert to have a complete bi-variate distribution $g_{<\theta_{PP},R>}$ for his prior beliefs. As we have discussed elsewhere (Bishop, Bloomfield et al. 2011), experts have great difficulty expressing beliefs about *dependence*, and even providing complete *marginal* prior distributions.

In what follows, then, we shall obtain results that require only *marginal* prior beliefs about the unknown model parameters, and these marginal beliefs will themselves be only *partial* – typically only one or two percentiles. Of course, reducing the burden on experts in these ways introduces further conservatism in the results (as in, for example, (Littlewood and Povyakalo 2013)).

Particularly problematic here is the parameter *R*. *R* is the probability that a randomly selected product from the development process is not perfect and passed *n* tests. We shall discuss later in the paper the difficulties of expressing prior beliefs about *R* (and about θ_{PP}); in the following sections, in order to develop the theory further, we shall take it that it *is* possible to express partial marginal prior beliefs about these two parameters.

[5] Strictly, since it is a function of *n*, it would be more precise to talk of *R*(*n*) here and in what follows. We shall use R instead just for notational simplicity.

4.2 Bayesian learning with one marginal percentile of θ_{PP} and R respectively

As in Section 3, we consider here first the case where an expert assessor is able to provide only very minimal prior beliefs about the unknown parameters. Specifically he is only prepared to provide the four numbers that constitute a single marginal percentile for each:

$$P(\theta_{PP} < y) = \gamma_\theta \tag{4.2.1}$$

$$P(R < r) = \gamma_r \tag{4.2.2}$$

To illustrate our approach, we restrict ourselves to what we consider to be the case most likely to crop up in practice, in which y+r⩽1. (The reasoning in other cases is similar, but with some tedious differences of detail.)

As we explained in Section 1, at the end of introductory remarks about our approach, although a standard Bayesian analysis would here require an expression of full prior beliefs about a *pair*, now, of parameters θ_{PP} and R, we continue instead to pursue our less ambitious "conservative" treatment. This involves working from only the *partial* elicitation of this bivariate prior represented by constraints (4.2.1) and (4.2.2). Although not fully elicited, our approach nevertheless reasons with the *concept* of the potential bi-variate prior distribution. We shall denote such a hypothetical bi-variate joint distribution, of θ_{PP} and R, by $g_{<\theta_{PP},R>}$ illustrated in Figure 8. Because of the constraint $\theta_{PP} + R \leq 1$, this distribution would have non-zero density only beneath the broken line connecting (0,1) and (1,0). In this triangle we label the probability masses associated with the four regions in the figure M_1, M_2, M_3, M_4. Then we have $P(\theta_{PP} < y) = \gamma_\theta = M_1 + M_3$ and $P(R < r) = \gamma_r = M_1 + M_2$.

To assist with interpretation of Figure 8, note that the quantity $1 - \theta_{PP} - R$ is the probability that a randomly selected product will fail one or more of the n tests applied to it.

In the figure this is simply the vertical (or horizontal[6]) distance of a point (R, θ_{PP}) of this distribution's support set below (respectively, to the left of) the dashed 45°-line.

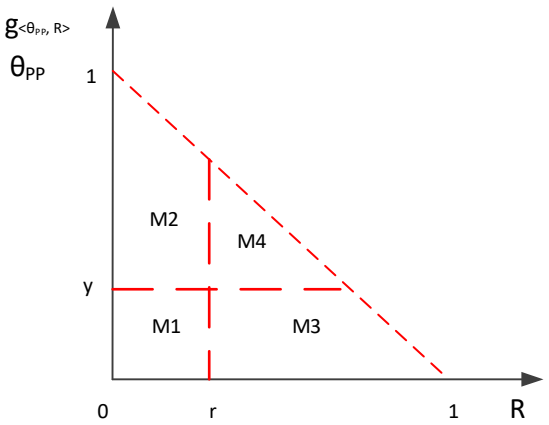

Figure 8 The prior constraints (4.2.1) and (4.2.2) on the joint distribution of $\boldsymbol{\theta_{PP}}$ and R

Our interest lies in the posterior confidence bound for θ_{PP}:

[6] The two distances are equal.

$$P(\theta_{PP} < y|\ process\ evidence)$$
$$= \frac{\int_0^1 \int_0^1 (I_{\theta_{PP}<y}(\theta_{PP}) \times [\theta_{PP} + R]^k) g_{<\theta_{PP},R>}(\theta_{PP}, R) d\theta_{PP} dR}{\int_0^1 \int_0^1 [\theta_{PP} + R]^k g_{<\theta_{PP},R>}(\theta_{PP}, R) d\theta_{PP} dR}$$

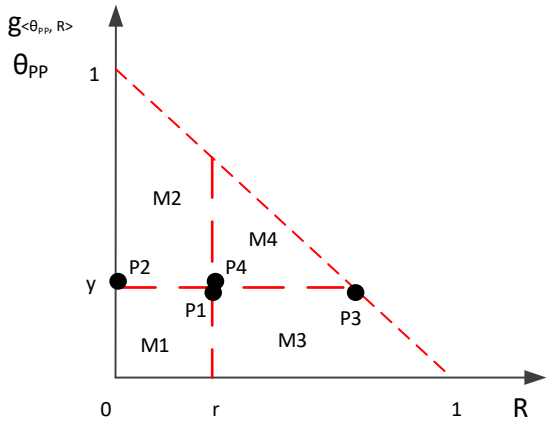

Figure 9 The most conservative joint prior distribution satisfying the constraints (4.2.1) and (4.2.2)

We can show (see Appendix 3) that the most conservative joint prior, g, satisfying the expert's constraints (4.2.1) and (4.2.2) is the 4-point-mass distribution shown in Figure 9, the black dots, P_1, P_2, P_3, P_4, representing the probability masses of the related value ranges. So that, for example, these four points' weighted mean distance (vertically or horizontally) from the dashed 45°-line is the prior subjective probability of the event that a randomly selected product will fail one or more of the first n tests applied to it, in the judgement of a hypothetical expert who happens to possess *exactly* this "most conservative" bivariate prior distribution. The corresponding most conservative posterior confidence bound for θ_{PP} is:

$$P(\theta_{PP} < y \,|\, \text{process evidence}) \le \frac{1}{1 + \dfrac{y^k \gamma_r + (y+r)^k (1 - \gamma_\theta - \gamma_r)}{\gamma_\theta}} = \gamma_\theta^* \qquad (4.2.3)$$

So γ_θ^* is the guaranteed-conservative posterior bound for the probability of perfection, θ_{PP}. Of the infinite number of prior distributions satisfying the expert's prior constraints (4.2.1) and (4.2.2), none result in a value for $P(\theta_{PP} < y|\ process\ evidence)$ smaller than this.

Table II shows some numerical examples.

Table II numerical examples of the "counter-intuitive" results

Case #	y	γ_θ	r	γ_r	k	γ_θ^*
1	0.9	0.05	0.09	0.05	10	0.056729362
2	0.9	0.05	0.099	0.05	10	0.052166239
3	0.99	0.05	0.009	0.05	10	0.050696597
4	0.99	0.05	0.0099	0.05	10	0.050285647

In fact, Table II shows some very counter-intuitive results: in all cases we find that $\gamma_\theta^* > \gamma_\theta$. That is, observing "good news" process evidence (failure-free working of *k* products) results in the posterior belief in perfection being *worse* than the prior belief.

In the next section we discuss the reasons for this result – essentially we show that the prior beliefs here are *too minimal* to be useful. This analysis suggests that priors be made sufficiently "partial" to make the expert's task feasible, but at the same time sufficiently informative to produce useful results.

4.3 The "counter-intuitive" results

The results above are "counter-intuitive" because seeing good evidence did not enhance our confidence in the positive claim about probability of perfection that interests us; on the contrary it increased our doubt. This result is not due to the choice of numbers in the examples of Table II: for any valid numbers of the model parameters the "counter-intuitive" result (i.e. $\gamma_\theta^* > \gamma_\theta$) will apply.

To understand this result, consider the most pessimistic prior distribution (satisfying the percentile constraints (4.2.1) and (4.2.2)), in Figure 9.

As shown in Appendix 3, this is a bi-variate distribution with probability mass concentrated on four points: P_1 at (r-, y-)[7], P_2 at (0, y), P_3 at (1-y, y-) and P_4 at (r, y). The corresponding masses at these points are M_1, M_2, M_3, M_4 respectively. The effect of seeing more and more process evidence is that in the posterior distribution the masses at P_2, P_1 and P_4 all gradually move to P_3. Since P_3(1-y, y-) is below the line θ_{PP}=y, it follows that

$$P(\theta_{PP} < y | \text{ process evidence}) > P(\theta_{PP} < y)$$

will always hold.

What does this mean practically? The four points of probability mass in Figure 9 can be thought of as representing four different development processes:

- The points P_1 and P_4 represent very similar processes that produce *some* (100y%) perfect, *some* (100r%) reliable (but not perfect) and *some* (100(1-y-r)%) unreliable products.
- The P_2 process produces *some* (100y%) perfect, *no* (0%) reliable (but not perfect) and *many* (100(1-y)%) unreliable products.
- The P_3 process produces *some* (100y%) perfect, *many* (100(1-y)%) reliable (but not perfect) and *no* (0%) unreliable products.

As we accumulate evidence of only good working, we tend to believe more strongly that our development process is P_3, i.e. the one with no unreliable products. That is, seeing good process evidence we tend to rule out processes producing unreliable products. And in our worst case, the only process producing no unreliable products is P_3. But P_3's ability to produce perfect products is lower than what we want to claim (as the ordinate is y- which is smaller than y). Therefore, believing more strongly in the process P_3 means having increasing doubt that the probability that randomly selected software from the process is not perfect is lower than a claimed bound.

But why are there only these 4 development processes as candidates here? Ideally, for example, there could be a development process producing many perfect products, based on

[7] Note the difference with (r, y): (r-, y-) is the coordinate infinitesimally close to (r, y) but smaller than r and y in both directions. The minus signs used here and later means smaller but infinitesimally close to the number.

the very best practices and design principles of utmost simplicity. If we had a prior belief that there exists this kind of process (even with very small probability), we would expect helpful support from the evidence to increase our confidence about the probability of perfection. Unfortunately, prior belief in the possibility of this kind of "perfection favoured" development process is ruled out due to the conservative nature of our method. That is, starting with the very minimal prior belief of only one marginal percentile of θ_{PP} and R respectively (i.e. constraints (4.2.1) and (4.2.2)), the most conservative joint prior distribution does not allow the possibility of "perfection favoured" process.

The counter-intuitive results, then, arise because in our pursuit of simplicity to aid the expert's task, we have allowed an expert to express partial prior beliefs that are *too* minimal. In what follows we suggest to relax this minimalism slightly: the cost, of course, is a heavier burden on the expert in expressing prior beliefs.

4.4 Using less minimal prior knowledge to get useful results

As discussed above, the most conservative prior distribution in Figure 9 did not include the possibility of a development process producing many perfect software products. Mathematically, this is due to the possibility for the marginal distribution of θ_{PP} to have zero variance (since only one prior percentile of θ_{PP} has been specified). Thus in the worst case, the marginal distribution of θ_{PP} will conservatively collapse onto one point (the *y* point of the vertical axis in Figure 9). Informally this collapsing means all our candidate development processes have the *similar* capability to produce perfect products. Therefore the assessor will believe that the good process evidence is *only* due to high reliability property of the process. Unfortunately, in our worst case (Figure 9), the one having the highest reliability property is one having slightly less capability of producing perfect products than our interest.

By extending our priors about θ_{PP} to at least two percentiles, we rule out the possibility of a prior having zero variance and so stop the collapsing onto a single point of the marginal distribution of θ_{PP}. So, we would have candidate processes with different capabilities of producing perfect products. Now perfection property might be thought to be the reason for seeing good process evidence. However, due to the conservative nature of our method, the "perfection favoured" process in our priors also tends to produce unreliable products, and the "high reliability (but not perfect) favoured" process will never produce unreliable products. So the "high reliability (but not perfect) favoured" process will always be preferred in the Bayesian learning. But this is obviously unrealistic, as no process will never produce unreliable products. So there must be an upper bound on *R* to make room for the possibility of unreliable products produced in all possible candidate processes.

We propose, therefore, a new (less) minimal partial prior belief about the model parameters in terms of two percentiles for θ_{PP} and one percentile and a certain upper bound for *R*:

$$P(\theta_{PP} < y_1) = \gamma_{\theta 1} \tag{4.4.1}$$

$$P(y_1 \leq \theta_{PP} < y_2) = \gamma_{\theta 2} \tag{4.4.2}$$

$$P(R < r) = \gamma_r \tag{4.4.3}$$

$$P(R < r_U) = 1 \tag{4.4.4}$$

The intention is that this set of prior beliefs may be sufficiently rich to produce useful results whilst still imposing upon an expert a manageable task in their expression. Figure 10 shows these prior beliefs, where the regions M_1, M_2,…M_6 contain non-zero probability mass. (Note that the position of the vertical line at r_U in Figure 10 is just for illustration; its precise location will be discussed in more detail later)

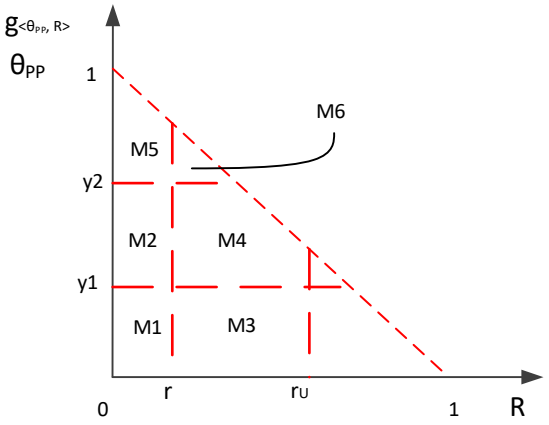

Figure 10 The minimum useful priori constraints on the joint distribution of $\boldsymbol{\theta_{PP}}$ and R

From the definition of R, there is the constraint $\theta_{PP} + R \leq 1$ which derives the other two implicit constraints $y_1 + r < 1$ and $y_2 + r < 1$.

The possible range of values of the certain upper bound r_U are: $1 - y_1 \leq r_U < 1$, $1 - y_2 \leq r_U < 1 - y_1$ and $r \leq r_U < 1 - y_2$. In Appendix 4 we show that useful results can be obtained only when the objective function is $P(\theta_{PP} < y_1 | \, process\ evidence)$[8] and r_U lies in the range $1 - y_2 \leq r_U < 1 - y_1$ or in the range $r \leq r_U < 1 - y_2$. The corresponding most pessimistic joint prior distributions are in figure 11 and figure 12 respectively. As before these are distributions with only a number of point masses.

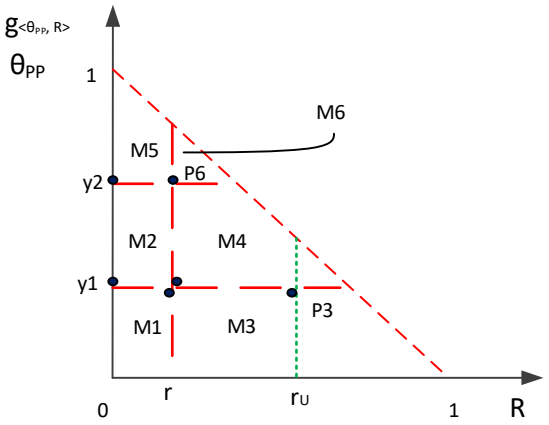

Figure 11 The most conservative prior satisfying the minimum useful priori constraints with $\boldsymbol{1 - y_2 \leq r_U < 1 - y_1}$

[8] As we have two prior confidence bounds on θ_{PP}, i.e. (4.4.1) and (4.4.2), both of them could be updated and potentially used as the input required by the LP model. But by proof, we found that only the former could be useful.

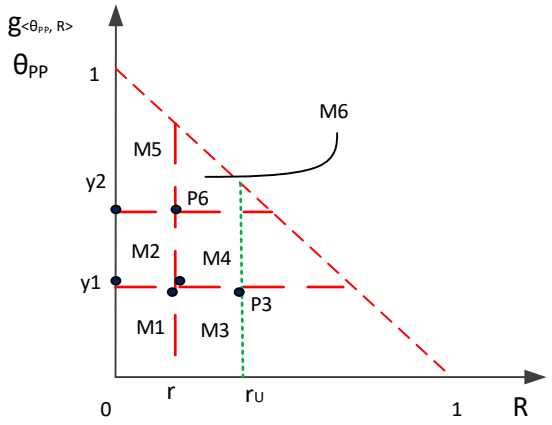

Figure 12 The most conservative prior satisfying the minimum useful priori constraints with $\boldsymbol{r \le r_U < 1 - y_2}$

With the most conservative prior distributions above, the objective function satisfies the formula below:

$$P(\theta_{PP} < y_1 | \text{ process evidence})$$
$$= \frac{\int_0^1 \int_0^1 \left(I_{\theta_{PP}<y1}(\theta_{PP}) \times [\theta_{PP} + R]^k\right) g_{<\theta_{PP},R>}(\theta_{PP}, R) d\theta_{PP} dR}{\int_0^1 \int_0^1 [\theta_{PP} + R]^k g_{<\theta_{PP},R>}(\theta_{PP}, R) d\theta_{PP} dR}$$
$$\le \frac{[y_1 + r_U]^k \gamma_{\theta 1}}{[y_1 + r_U]^k \gamma_{\theta 1} + [y_1 + r]^k \gamma_{\theta 2} + {y_2}^k \gamma_r + [y_2 + r]^k (1 - \gamma_r - \gamma_{\theta 2} - \gamma_{\theta 1})} = \gamma^*_{\theta 1} \quad (4.4.5)$$

That is, $\gamma^*_{\theta 1}$ is the most conservative posterior belief we are interested in.

Thus, for example, the pair $(1 - y_1, \gamma^*_{\theta 1})$ could be used as the input pair (p_B, α_B) of the LP model (Littlewood and Povyakalo 2013) as in equations (1.2) and (1.3) in Section 1.

Table III shows some numerical examples. The first 4 cases are illustrating the results with constraint $1 - y_2 \le r_U < 1 - y_1$ and the rest of them are about the ones with constraint $r \le r_U < 1 - y_2$.

Table III Numerical examples using the "less minimal" prior knowledge

#	y_1	$\gamma_{\theta 1}$	y_2	$\gamma_{\theta 2}$	r	γ_r	r_U	k	$\gamma^*_{\theta 1}$
1	0.9	0.05	0.99	0.05	0.009	0.05	0.0600	10	0.035391421
2	0.9	0.05	0.99	0.05	0.009	0.05	0.0999	10	0.052250881
3	0.9	0.05	0.99	0.05	0.009	0.05	0.0110	10	0.021265865
4	0.9	0.05	0.99	0.05	0.009	0.05	0.0600	50	0.007679250
5	0.9	0.05	0.99	0.05	0.009	0.05	0.0095	10	0.020925570
6	0.9	0.05	0.92	0.05	0.009	0.05	0.0790	10	0.082798474
7	0.9	0.05	0.99	0.05	0.009	0.05	0.0091	10	0.020835634
8	0.9	0.05	0.99	0.05	0.009	0.05	0.0095	50	0.000518818

The $\gamma^*_{\theta 1}$ column in Table III contains results that are both bigger and smaller than the prior value, $\gamma_{\theta 1} = 0.05$. It seems the "less minimal" prior constraints are indeed helpful in some cases compared with the counter-intuitive results in the section 4.3, but not universally.

The two counter-intuitive results happen in cases #2 and #6 in which the certain upper bound r_U is high. Mathematically, there is a competition between the points P_3 and P_6 in Figures 11 and 12, in which the result depends on the specific choices of the parameters. The higher the certain upper bound r_U, the more likely P_3 will be the result. The informal interpretation in that case is that we believe more in the P_3 process which is producing many "ultra-reliable but not perfect" products.

In contrast, when the result is the P_6 process we have our "perfection favoured" process.

Cases #4 and #8 get practically useful results which are one and two orders of magnitude better that prior belief, respectively. Comparing with case 1 and 5 respectively where only the numbers of *k* vary, there is evidence that, for a certain set of prior numbers, collecting more products in the process evidence is very helpful. This is consistent with the conclusion in section 3.2, and is in accord with intuition. But our comments about the feasibility of observing many previous products still apply.

Notice that, in contrast to *k*, we cannot say how the parameter *n* affects the result, since *n* is now hidden in the subjective beliefs about *R*. That said, from the definition of *R*, it is obvious that larger *n* means smaller *R*.

Comparing case #3 to case #1 and case #7 to case #5, the smaller values of r_U here indeed give better results; but its impact does not seem as great as that of *k*.

4.5 How to use these results? A possible model of negotiation

We imagine a situation where there is to be negotiation between a regulator and the licensee of a system – say a nuclear plant – that contains a critical software-based subsystem. As part of the safety case for the plant, the licensee needs to make a claim for "probable perfection" of this subsystem. That is, he needs to convince the regulator to accept as reasonable his declared confidence that θ_{PP} is no smaller than some declared number y_1.

Clearly this cannot be done by mere assertion about the licensee's top-level claim here. It is agreed between the regulator and the licensee that their negotiation will take place within the framework of our approach outlined above. Our model then allows the discussion between them to be a negotiation about the elements of the argument that underpin the regulator's claim: i.e. about numbers for the 7 parameters $< y_1, \gamma_{\theta 1}, y_2, \gamma_{\theta 2}, r, \gamma_r, r_U >$ needed to obtain the conservative posterior bounding confidence, (4.4.5). Recall that there are some constraints on the numerical values that these numbers need to satisfy: these constraints must be respected in the discussion.

Rather than say simply "trust me when I declare my confidence in perfection", the licensee says "here are the numbers I used to arrive at my confidence in perfection". *These numbers*, then, rather than simply the top-level claim, become the subject of discussion and negotiation with the regulator.

Below we sketch briefly how such negotiation might be conducted.

We begin with a simple monotonicity analysis of the impact of the seven parameters on $\gamma^*_{\theta 1}$. It is easy to see (by partial differentiation w.r.t. each parameter) that:

1. $\gamma^*_{\theta 1}$ is an increasing function in terms of $\gamma_{\theta 1}$.
2. $\gamma^*_{\theta 1}$ is a decreasing function in terms of y_2.

3. $\gamma_{\theta 1}^{*}$ is an increasing function in terms of $\gamma_{\theta 2}$.
4. $\gamma_{\theta 1}^{*}$ is a decreasing function in terms of r.
5. $\gamma_{\theta 1}^{*}$ is an increasing function in terms of γ_{r}.
6. $\gamma_{\theta 1}^{*}$ is an increasing function in terms of r_U.
7. The monotonicity of y_1 depends on specific choices of other parameters

The monotonicity analysis shows how changes to these parameters make claims about probability of perfection stronger or weaker: increasing parameters in 1, 3, 5, 6 make the posterior bound more conservative.

Note that 7 parameters are either "bounds" (y_1, y_2, r, r_U) or their corresponding "confidences" ($\gamma_{\theta 1}, \gamma_{\theta 2}, \gamma_r$) which represent the licensee beliefs. It seems likely that agreement on the bounds will be easier, and negotiation will then concentrate on the related confidences. The bound y_1 is a special bound which is essentially the licensee's claim, his confidence in which he is trying to persuade the regulator is acceptable.

The negotiation could proceed in the following steps:

- They begin by agreeing on the "claim", y_1: this comes from higher-level system requirements. Then the licensee and regulator state their prior beliefs about this claim, i.e. values for $\gamma_{\theta 1}$.
- Second, licensee and regulator agree on y_2 and r. Then they state their beliefs about y_2 and r: i.e. values for $\gamma_{\theta 2}$ and γ_r.
- Finally, they negotiate on the r_U.

The following is a simple example of how this might proceed (the numbers are merely illustrative and not meant to be representative of those that would occur in a real negotiation about a critical system):

1. The claim concerning probability of perfection is *pnp*<0.1 (obtained from higher level requirements), so the y_1=0.9. For $\gamma_{\theta 1}$, the licensee's doubt that the probability of perfection is bigger than 0.9 is 0.05, and the regulator's doubt about it is 0.1. Then $0.05 \leq \gamma_{\theta 1} \leq 0.1$.
2. Then the two parties fix the $y_2 = 0.99$ and $r = 0.009$
    a. For $\gamma_{\theta 2}$, the licensee's doubt that the θ_{pp} is bigger than 0.99 is 0.1, and the regulator's doubt about it is 0.2. Then $0.05 \leq \gamma_{\theta 2} \leq 0.1$.
    b. For γ_r, the licensee's doubt that the *R* is bigger than 0.009 is 0.05, and the regulator's doubt about it is 0.1. Then $0.05 \leq \gamma_r \leq 0.1$.
3. The regulator does not believe that the R will be bigger than 0.0098, and the licensee does not believe that R will be bigger than 0.0095. Then $0.0095 \leq r_u \leq 0.0098$.

For each confidence/doubt here we have assumed that the regulator will be more conservative in his beliefs than the licensee. If the process evidence to form posterior beliefs is based on *k*=50 products, we have the following Table IV:

Table IV An illustrative numerical example of the negotiation model

y_1	$\gamma_{\theta 1}$	y_2	$\gamma_{\theta 2}$	r	γ_r	r_U	k	best $\gamma^*_{\theta 1}$	worst $\gamma^*_{\theta 1}$
0.9	[0.05, 0.1]	0.99	[0.05, 0.1]	0.009	[0.05, 0.1]	[0.0095, 0.0098]	50	0.00052	0.0012

There is a factor of about 2 difference in the doubt of the regulator ("worst" in the Table) and the licensee about the claim of perfection.

The reduction in doubt about probability of perfection bound (i.e. from prior to posterior) is about the same for licensee and licensor in this example: around two orders of magnitude in each case.

In the next example, Table V, there is significant variation (about an order of magnitude) between the two parties on parameter r_U, keeping the other parameters the same as before.

Table V A numerical example of the negotiation model with big variance on r_U.

y_1	$\gamma_{\theta 1}$	y_2	$\gamma_{\theta 2}$	r	γ_r	r_U	k	best $\gamma^*_{\theta 1}$	worst $\gamma^*_{\theta 1}$
0.9	[0.05, 0.1]	0.99	[0.05, 0.1]	0.009	[0.05, 0.1]	[0.0091, 0.098]	50	0.00051	0.110

Introducing this significant difference in the parties' views about r_U results in the best and the worst result being very different (the worst result even shows the "counter-intuitive" result). This suggests that the value of r_U is critical in negotiation.

Table VI shows some more results with different parameter values for each party. From this limited analysis it seems that prior belief in $\gamma_{\theta 1}$ also has a high impact on the final result. In contrast, wide variation in γ_r and $\gamma_{\theta 2}$ seemed to have less impact.

From the limited evidence of these examples, it seems negotiation might centre upon $\gamma_{\theta 1}$ and r_U, since these seem to have the greatest effect on the result $\gamma^*_{\theta 1}$.

The examples here are meant to be only illustrative, so it would be wrong to draw strong conclusions from the particular numerical results above – we have not attempted to make these numbers typical of real systems. Rather our intention is to show how our modeling might provide a framework for negotiation. It does this by allowing the discussion to take place about the *components* of the parties' arguments that support their different claims about perfection. We believe that, of the 7 parameters in total in the model, it is likely that the parties' differences will centre upon 4 of these parameters: the 3 parameters representing *confidence*, i.e. $\gamma_{\theta 1}$, $\gamma_{\theta 2}$, γ_r, together with r_u.

Table VI Numerical examples of the negotiation model with different variance on the parameters

y_1	$\gamma_{\theta 1}$	y_2	$\gamma_{\theta 2}$	r	γ_r	r_U	k	best $\gamma^*_{\theta 1}$	worst $\gamma^*_{\theta 1}$
0.9	[0.05, 0.1]	0.99	[0.05, 0.1]	0.009	[0.05, 0.1]	[0.0095, 0.0098]	50	0.00052	0.0012
0.9	[0.05, 0.1]	0.99	[0.05, 0.1]	0.009	[0.05, 0.5]	[0.0095, 0.0098]	50	0.00052	0.0015
0.9	[0.05,0.1]	0.99	[0.05, 0.5]	0.009	[0.05, 0.1]	[0.0095, 0.0098]	50	0.00052	0.0024
0.9	[0.05, 0.5]	0.99	[0.05, 0.1]	0.009	[0.05, 0.1]	[0.0095, 0.0098]	50	0.00052	0.011

5. Conclusions and discussion

The model we have presented here allows evidence of the efficacy of development process(es) to be taken into account when assessing the possibility of perfection of a software-based system. Informal arguments in support of system dependability have long used the following kind of evidence: "we have in place a wealth of experience and good processes, as evidenced by the many similar systems we have built in the past that have experienced lots of operational exposure without failure – this makes us confident that *this* system will operate with high dependability." Such claims about track records are, of course, intuitively attractive: e.g. we are confident flying in a new aircraft type built by Boeing or Airbus because we have seen the excellent safety record of previous types. The work reported here is an attempt to put this kind of argument onto a formal basis. In particular, it allows *quantitative claims* to be made – in this case, about possible perfection – and it does this in ways that are *guaranteed to be conservative*.

Our mathematical approach to the problem is via a *hierarchical model* (Gelman, Carlin et al. 2013) – in fact a two-stage Bayesian model. Such models are in wide use, and in particular have been used in the nuclear industry (Kaplan 1983, Bunea, Charitos et al. 2005), which was the source of our own initial interest in these issues. In conventional applications of such two-stage Bayesian models, particular attention must be paid to the assumptions about the hyperdistribution and the hyperpriors for the parameters of that distribution (Cooke, Bunea et al. 2002, Vaurio and Jänkälä 2006). Our approach in this paper is different: for our two-stage Bayesian model we assume an *arbitrary* hyperdistribution, and obtain worst-case priors for its parameters. We believe this approach is new.

In the nuclear examples in the papers cited above, failure data is assimilated from different plants and/or their components, and the validity of the models will depend on the similarity of usage across this population of users. The database ZEBD (PowerTech 2010) provides datasets for various two-stage Bayesian models of components used in German nuclear power plants. Similarly, our modeling here depends upon the reasonableness of notions of "similarity" between different products when these have been developed using the same process, and the way that these are represented mathematically.

In this section we discuss briefly some of the issues that arise in using our model to make claims about perfection of real systems. As the attentive reader will have noticed, an expert assessor would face some difficult problems in using the model. We discuss these, and point to ways forward that need to be addressed in further work.

5.1 Similarity and common development process

We have used informally terms like "similarity" and "common development process". What would these mean in practice? In other words, when could one claim that one has identified a population of similar products, of which the one being assessed is an instance? The two-stage model requires that the products whose record of success we use as evidence, and the product to be assessed, all be sampled independently from the same distribution.

This matter of "similarity" is of course the crucial issue in any application of statistics to predict something about one individual. It usually needs to be argued informally on the basis of what is known about the population and whether the individual can be considered as being sampled randomly from it.

For software products, such a claim could be quite convincing, for instance, if it came from a company that has developed multiple products in a "product line", using an identical "software engineering process" (set of methods and tools for development and verification, artefacts produced and their required verification and documentation, etc.), the same team, under similar conditions (e.g., how budget and deadlines are set). The variations of achieved dependability (a product's *pfd*, and in particular whether it is perfect or not) between these products would be just contributed by those accidental events in the application of the process that cannot be systematically observed and are thus treated as random factors. We would expect this scenario to give a comparatively "concentrated" distribution for the parameters, a "consistent" process, as we called it in section 2, although this is not a formal requirement for the model to apply.

In principle, one may wish to apply a broader definition of "similarity" – e.g., all the products in a certain range of code size, developed for a certain class of applications, according to the practices prescribed by a certain standard for a certain criticality level, by companies within a certain range of CMM "maturity levels". But a claim based on a sample of such a more diverse population would be harder to believe, because it would be harder to believe that all the assumptions are satisfied: that a little more effort would not show some systematic differences between these products (if e.g. we knew that the producer of this one product has better records – or worse – than the bulk of the industry; or that the application problems addressed can really be classified into subsets for which we would expect different difficulties in achieving good *pfd* and perfection*;* then not using this information about the new product to be assessed would ignore useful information); that indeed the claims that all have passed *n* demands are equally believable for all the products (for some we might not really trust the failure reporting process); that the set of *k* previous products chosen is a true random sample of the population and not affected by some involuntary selection bias (we are not really informed about *all* products in this population: are we accidentally picking our sample based on some factor that correlated with their true reliability?).

5.2 Extent of process evidence: what values of n and k are feasible?

It is trivially true that the more products have performed successfully (the higher *k* is), and the more failure-free demands have been seen for them (the higher *n* is), the better. However, our numerical results here, whilst only intended to be illustrative rather than realistic, suggest – not surprisingly – that *k* seems more important than *n*. Massive operational exposure from only a very few previous products tells us less about the population of products than more modest exposure of many products.

Unfortunately, the value of *k* is unlikely to be controllable. It is infeasible to build many products just to evaluate the efficacy of a development process. Instead, users of a model like this must rely upon the evidence that is available. The scenario of a "product line" mentioned above is a case in point. Likewise, in a company that has built several generations of safety protection systems, there may be evidence of their reliability in operational use that could be used for our model. But values of *k* in such cases are likely to be modest, unlikely to be as large as the value $k = 50$ in our earlier tables, which gave the most useful results.

5.3 Inference from deployment of the same product in many environments

Although we have described everything here in terms of *multiple similar products* developed using the same process (to tackle similar problems), the same model can be applied to the case where a *single product* is used in *multiple similar* (but different) *environments*.

For example, some components of safety protection systems are installed in many industrial plants. Each plant has a different operational environment (probability distribution of the demand events that may happen - timing, values of inputs etc.), leading in principle to a different *pfd* in each installation (or zero *pfd*, if the component is "perfect"). If such a component is reused in a new plant for which it can be argued that its environment is just one more instance of the same general population, then our two-stage model can be used, and *k* may be quite large, allowing reasonably strong conclusions about probability of perfection.

We note, however, a limitation here. In Section 1 we gave two reasons for our interest in possible perfection: firstly that it is of value in its own right for making claims about lifetime reliability; secondly that it overcomes a basic hurdle in making reliability claims for fault-tolerant 1-out-of-2 systems via the result (1.1) from (Littlewood and Rushby 2012). This new interpretation of similarity does not fit into the latter scenario. The reason is as follows.

The basic result in (Littlewood and Rushby 2012), which a reader may think we could use to obtain our equation (1.1) is, from a slightly modified form of equation (6) on p 1182 of (Littlewood and Rushby 2012):

$$P(\text{system fails a random demand} \mid \text{this operational environment}, pfd_A, pnp_B) \leq$$

$$P(A \text{ fails a random demand}, B \text{ is imperfect} \mid \text{this environment}, pfd_A, pnp_B) =$$

$$P(A \text{ fails a random demand} \mid \text{this environment}, pfd_A, pnp_B) \times$$

$$P(B \text{ is imperfect} \mid \text{this environment}, pfd_A, pnp_B)$$

Now the first term on the RHS is

$$P(A \text{ fails a random demand} \mid \text{this environment}, pfd_A, pnp_B) = pfd_A$$

The second term on the RHS is $P(B \text{ is imperfect} \mid \text{this environment}, pfd_A, pnp_B)$.

Unfortunately, in contrast to the theorem on p1182 of (Littlewood and Rushby 2012), this term is not pnp_B. Instead, the pnp_B here is the probability that B is imperfect in a *randomly chosen environment,* not in *this* environment (i.e. the one from which pfd_A is defined as the probability of failure of a randomly selected demand from it). We therefore cannot claim that the system *pfd* is bounded by the simple product of pfd_A and pnp_B as required for our (1.1).

There is a practical restriction on the allowable values of *n*: all our results assume that *n* takes the same value over all *k* systems observed. Clearly, this is a serious issue because it is unlikely to be true in practice. It arises because the parameter *R* involves a particular *n*. It would be infeasible to elicit beliefs about *k* different R_i corresponding to the different values of n_i for the *k* different products. A work-around is to choose conservatively the minimum n_i to be the *n* in our equations. Of course, this solution may be very conservative if the minimum n_i is much smaller than other n_i. In this case, we could simply ignore the evidence from this product, and carry out the model calculations using the next smallest n_i. The price, of course, is a reduction of *k* to *k*-1.

5.4 What is *R*, and how could an expert express beliefs about it?

Our general model is complex, and at its heart lies a distribution for *pfd* which we have called *f*: see the middle level of Figure 7. Specifying this distribution is difficult. As we argued earlier, adopting some parametric family for *f* and then incorporating its unknown parameter values into the Bayesian analysis does not seem a plausible way forward. It would be hard to

justify a particular choice of family; given such a family, it would be unreasonable to expect an expert to be able to express even partial beliefs about its parameters.

The result of Section 4.1 is a way round these difficulties, albeit via the introduction of different ones. The key result here shows that knowledge of the complete (non-zero) shape of *f* does not need to be specified, because knowledge of *R* is "sufficient" (with θ_{pp}), in a precise mathematical sense, for statistical inference about perfection. We commented that this notion of sufficiency is similar to the idea of "sufficient statistic" in classical statistics.

Whilst the use of *R* is a considerable simplification, compared with the need to specify a complete distribution *f*, it requires an expert to express (at least partial) prior beliefs about *R*. How could he do this?

From Section 4.1, $R = \int_{0+}^{1} (1-p)^n f(p|\theta)\, dp$. Here $f(p|\theta)$ is the distribution of the *pfd* of a randomly selected software system from the population produced by the common development process for this and similar problems. Note that the range of *p* excludes the perfect products, so *R* means *the probability that a randomly selected product from the population produced by the development process is not perfect but passes* n *tests.*

One way to think about *R* is to imagine some extremely long sequence of randomly selected tests and to say that *R* is the probability that the system will fail on at least one of these tests, but that this failure will not occur on the first *n* tests applied. It can thus be seen as involving belief about both the *fact* of imperfection of the system, and the "size" of that imperfection.

As a shorthand, we shall say that *R* means "reliable and not perfect"; where by "reliable" we mean "passes *n* tests without failure". Note the presence of *n* in this: in fact in this treatment of our model, *n* only appears in the Bayesian analysis via an expert's belief(s) about *R*.

Note also that *R*, like the probability of perfection θ_{PP}, is an *objective property* of the population of software systems: it is an unknown "in-the-world" parameter. There will therefore be epistemic uncertainty for assessors about these two parameters – in fact this is the only epistemic uncertainty in the model now. The task of the assessor is to express his (limited, partial) beliefs, for example in constraints like (4.2.2). How should he go about this? – this task seems harder for *R* than the similar task he faces concerning θ_{PP}.

Of course, scrupulous companies may be expected to have extensive (even complete) operational failure data on their earlier products. In some cases a product may have had massive exposure. Some will have experienced failures, some not. This is the evidence of previous experience of similar products upon which assessors must make their judgments about the parameters of the model.

So, in our shorthand terminology, we have

$$R = P(\text{imperfect and reliable}) = P(\text{imperfect})P(\text{reliable}|\text{imperfect}) \qquad (5.1.1)$$

$P(\text{imperfect})$ here is just $1 - \theta_{PP}$; we assume the assessor is able to express beliefs about θ_{PP}. More difficult is the second factor on the right hand side of (5.1.1), which is a conditional probability. It is well known people find it hard to assess a conditional probability – in fact it is essentially the same, and thus as difficult, as assessing bivariate uncertainty.

Some help may come here from noting that

$$P(\text{reliable}|\text{imperfect}) \leq P(\text{reliable})$$

from which it follows that

$$R = P(\text{reliable}) - \theta_{PP} \leq (1 - \theta_{PP})P(\text{reliable})$$

and the problem reduces to expressing beliefs about a product of unconditional probabilities, with all the difficulties related to epistemic dependence between them.

Clearly this remains a difficult problem, though, and is an obvious candidate for further thought.

5.5 Why conservative Bayesian reasoning?

In this and other recent work on systems dependability we have adopted a Bayesian approach to the treatment of uncertainty. A probabilistic treatment of risk is now standard in most safety-critical industrial domains; the Bayesian approach is widely and increasingly accepted: e.g. directives for Probabilistic Risk Assessment (PRA) like (NRC 2003) and (NRC 2017) recommend Bayesian methods. Comparing the 2017 version (NRC 2017) with its 2009 version (with No. UREG1855-V.1.2009), we see that the earlier version mentioned frequentist statistics as acceptable for parameter estimation, while pointing at its limits, but the 2017 version no longer does so. However, researchers and practitioners acknowledge as the main practical difficulty of Bayesian methods the difficulty of eliciting prior beliefs from experts who "own the problem" – in the terminology of this paper, regulators and licensees – in the kinds of complex situations we are dealing with. It is this issue that we have addressed here and in our earlier work.

In the Bayesian literature, including regulatory documents and advice for PRA practitioners, various approaches are proposed to address this problem: e.g. empirical Bayes, non-informative priors. For many application domains, it is possible to obtain very extensive data, and the differences between such approaches and "proper" Bayes, involving informative priors, disappear. Unfortunately, this is a luxury we do not have for the kinds of safety-critical applications we are considering. Indeed, in software engineering generally, large sample sizes are rare: e.g. random samples of *programs* are unknown, except in specialized small-scale experiments such as those described in (Knight and Leveson 1986, Eckhardt, Caglayan et al. 1991). In conclusion, the conditions for obtaining complete probability distributions from empirical data are not satisfied.

The method we have adopted in the face of these problems is, we believe, novel. It allows the expert to populate the model with minimal partial prior beliefs. Our procedure then is to obtain results for the claims of interest – in this case, confidence bounds for probability of perfection – that are guaranteed to be conservative. The trick here is to allow the user's prior beliefs to be as minimal as possible to aid his task, whilst not being *too* minimal to provide useful results. We gave examples of prior beliefs that *are* too minimal to be useful, and showed that prior beliefs that were only *slightly* more informative (i.e. that impose upon the assessor a task that is only *slightly* more onerous) could give useful results.

Whilst the model here is a complex one, we presented an example of how it might be used to inform the negotiation between a regulator and a licensee about a claim in part of a safety case. We claim, somewhat tentatively, that this shows the potential usefulness of our approach.

There are, of course, other proposed ways of addressing these difficulties. Interesting surveys and discussions of these alternatives are for instance (Aven and Zio 2011) and (Zio and Pedroni 2013). They characterize their application domain as "high-consequence risk analysis of complex systems with limited knowledge on their behavior." It is the analysis of such systems – particularly those in the nuclear industry – that motivated our own work. The alternatives they discuss include approaches like possibility theory and Dempster-Shafer evidence theory. These approaches differ somewhat radically from standard probability

calculus (although links can be made between the formalisms). They are not candidate solutions as of now, simply because the very axioms are controversial, and in any case adopting them would require a cultural change in the whole large community of practitioners. Other alternatives simply attempt to extend probability calculus by considering intervals of uncertainty on probability assignments. Our approach is more akin to these, but we require only minimal additions to standard Bayesian methods, via our theorems that state bounds for whole sets of prior distributions. This way we allow users of the method to see the true consequences of the minimal beliefs they state. An important aspect is that the theorems bound the results implied by whole sets of possible distributions, as opposed to the effects of just intervals of variation of parameters of parametric distributions. Currently applied sensitivity analysis can do the latter, but the very use of parametric distributions may sometimes impose the analysts' description of the problem on top of the experts' input to the analysis, which we agree with (Zio and Pedroni 2013, p.41) should be avoided.

5.6 Counter-intuitive results

We briefly comment on the counter-intuitive results of Sections 3 and 4. We see these as a warning against placing too much trust in informal reasoning when dealing with these quite complex models for dependability. We observed here cases where "obviously good news" from operational testing of a system could – counter-intuitively – decrease "confidence in perfection".

Similar counter-intuitive results were found in (Littlewood and Wright 2007). There, a two-legged argument was used to support claims for the *pfd* of a system. The two legs used were, respectively, evidence of failure-free working on test (involving a possibly fallible test oracle), and evidence of proof of correctness against a (possibly incorrect) formal specification. The counter-intuitive result there was that *more* failure-free working could, in certain circumstances, result in *lower confidence* in a small *pfd.*

In all these cases, of course, further analysis showed that the interpretation from the formal modeling was correct – i.e. eventually, intuition guided by formal reasoning caught up with reality. Such examples are, we believe, warnings against the use of unaided informal engineering judgment.

5.7 Some comments on applicability and practicality

Hierarchical models of the kind we have described here can be complex, and consequently impose upon users quite stringent requirements in a Bayesian analysis, particularly in expressing the necessary prior beliefs. The practical situations we are dealing with are not ones where there are large quantities of data, in which the dominant contribution to posterior belief via Bayes Theorem comes from the likelihood function. There are no believable "ignorance" priors. It follows that these are not situations in which "the data can speak for themselves."

Our approach to this problem in our recent work (Bishop, Bloomfield et al. 2011, Zhao 2015, Zhao 2017) has been to allow assessors to express only quite minimal prior beliefs (e.g. one or two percentiles, rather than complete distributional information), and then to obtain guaranteed-conservative posterior results under this minimal information set-up. We believe this way of handling the problem of prior belief is novel, and we have adopted it again in the work reported here.

Readers may think our proposed procedures are still rather complex. In fact an anonymous reviewer made the following remarks (we have edited them slightly):

"My major remaining concern is with the applicability of the results… The authors acknowledge this issue... I think it would be inappropriate to expect them to solve it in this paper, which establishes the intellectual framework for a certain new kind of assessment. After all, fifty years after the introduction of Hoare logic there are software developers claiming it is infeasible to use, as well as other developers using it routinely. The main reason for the charge of impracticality here can be regarded as primarily social, lying in the technical experience required of software developers, which is a manifestly different issue from the intellectual applicability of the method itself. Similarly, here, practicality is likely to follow possibility by a good few years. It is not necessarily required for assessors to be able to understand the framework; only for a regulator to do so, who then demands of assessors that they provide certain values of parameters..."

We think this is a fair assessment of the difficulties in using our results. Assessors will need to provide the 7 parameters contained in the key equations (4.4.1) to (4.4.4). However, as we suggest in our account of a possible negotiation regime in Section 4.5, it is likely that agreement will need to be negotiated on the values of only 4 of these: that is, the parameters y_1, y_2, and r should be the subject of early agreement, leaving only $\gamma_{\theta 1}$, $\gamma_{\theta 2}$, γ_r, and r_U for negotiation.

Whilst agreeing with sceptical readers that all this is still not easy, we nevertheless claim that it is a very much simpler task than expressing complete distributional prior beliefs. We also claim that it is a *minimal* simplification: as we have shown, greater simplification (e.g. requiring only one percentile rather than two) does not produce useful results. For anyone wishing to use this particular Bayesian hierarchical model, then, there is a sense in which our approach to its analysis is *as simple as possible, but no simpler*.

Having said that, our results are conservative. For the kinds of safety-critical systems that prompted our interest in this kind of modeling, such a guarantee of conservatism seems right. Depending on circumstances, though, the results may be disappointingly conservative for those wishing to be allowed to operate the systems under analysis. We see no way to avoid this.

Appendix 1

By the mean value theorem for integrals, we could find two values, say P_1 and P_2, satisfying the equations below:

$$L(P_1)\int_0^{y-} g(\theta_{PP})d\theta_{PP} = \int_0^{y-} L(\theta_{PP})g(\theta_{PP})d\theta_{PP}$$

$$L(P_2)\int_y^1 g(\theta_{PP})d\theta_{PP} = \int_y^1 L(\theta_{PP})g(\theta_{PP})d\theta_{PP}$$

where the L is the likelihood function as the formula (3.2) and $0 \le P_1 < y, y \le P_2 \le 1$. From the one percentile prior knowledge $\mathrm{P}(\theta_{PP} < \mathrm{y}) = \alpha_\theta$, we know:

$$\int_0^{y-} g(\theta_{PP})d\theta_{PP} = \alpha_\theta$$

$$\int_y^1 g(\theta_{PP})d\theta_{PP} = 1 - \alpha_\theta$$

Therefore our objective function:

$$\alpha_\theta{}^* = \frac{\int_0^y L(\theta_{PP})g(\theta_{PP})d\theta_{PP}}{\int_0^1 L(\theta_{PP})g(\theta_{PP})\,d\theta_{PP}} = \frac{L(P_1)*\alpha_\theta}{L(P_1)*\alpha_\theta + L(P_2)*(1-\alpha_\theta)} = \frac{1}{1+\frac{L(P_2)*(1-\alpha_\theta)}{L(P_1)*\alpha_\theta}}$$

As the likelihood function L is an increasing function:

$$\alpha_\theta{}^* = \frac{1}{1+\frac{L(P_2)*(1-\alpha_\theta)}{L(P_1)*\alpha_\theta}} \le \frac{1}{1+\frac{L(y)*(1-\alpha_\theta)}{L(y)*\alpha_\theta}} = \alpha_\theta$$

This means, with the most conservative prior $g(\theta_{PP})$ showed in figure 4, we cannot learn anything about probability of perfection from the process evidence.

Appendix 2

Similarly as the proof in Appendix 1 here, by the mean value theorem for integrals, we could find 3 values, say P_1, P_2 and P_3 satisfying the equations below:

$$L(P_1)\int_0^{y_1-} g(\theta_{PP})d\theta_{PP} = \int_0^{y_1-} L(\theta_{PP})g(\theta_{PP})d\theta_{PP}$$

$$L(P_2)\int_{y_1}^{y_2-} g(\theta_{PP})d\theta_{PP} = \int_{y_1}^{y_2-} L(\theta_{PP})g(\theta_{PP})d\theta_{PP}$$

$$L(P_3)\int_{y_2}^{1} g(\theta_{PP})d\theta_{PP} = \int_{y_2}^{1} L(\theta_{PP})g(\theta_{PP})d\theta_{PP}$$

where the L is the likelihood function as the formula (3.2). And

$$0 \le P_1 < y_1, y_1 \le P_2 < y_2, y_2 \le P_3 \le 1$$

From the two percentile prior knowledge $P(\theta_{PP} < y_1) = \alpha_{\theta 1}$ and $P(y_1 \le \theta_{PP} < y_2) = \alpha_{\theta 2}$, we know:

$$\int_0^{y_1-} g(\theta_{PP})d\theta_{PP} = \alpha_{\theta 1}$$

$$\int_{y_1}^{y_2-} g(\theta_{PP})d\theta_{PP} = \alpha_{\theta 2}$$

$$\int_{y_2}^{1} g(\theta_{PP})d\theta_{PP} = 1-\alpha_{\theta 1}-\alpha_{\theta 2}$$

Therefore our two objective functions:

$$\alpha^*_{\theta 1} = \frac{\int_0^{y_1-} L(\theta_{PP})g(\theta_{PP})d\theta_{PP}}{\int_0^1 L(\theta_{PP})g(\theta_{PP})\,d\theta_{PP}} = \frac{L(P_1)*\alpha_{\theta 1}}{L(P_1)*\alpha_{\theta 1} + L(P_2)*\alpha_{\theta 2} + L(P_3)*(1-\alpha_{\theta 1}-\alpha_{\theta 2})}$$

$$= \frac{1}{1+\frac{L(P_2)*\alpha_{\theta 2} + L(P_3)*(1-\alpha_{\theta 1}-\alpha_{\theta 2})}{L(P_1)*\alpha_{\theta 1}}}$$

$$\alpha^*_{\theta 1+\theta 2} = \frac{\int_0^{y_2^-} L(\theta_{PP}) g(\theta_{PP}) d\theta_{PP}}{\int_0^1 L(\theta_{PP}) g(\theta_{PP})\, d\theta_{PP}} = \frac{L(P_1) * \alpha_{\theta 1} + L(P_2) * \alpha_{\theta 2}}{L(P_1) * \alpha_{\theta 1} + L(P_2) * \alpha_{\theta 2} + L(P_3) * (1 - \alpha_{\theta 1} - \alpha_{\theta 2})}$$
$$= \frac{1}{1 + \frac{L(P_3) * (1 - \alpha_{\theta 1} - \alpha_{\theta 2})}{L(P_1) * \alpha_{\theta 1} + L(P_2) * \alpha_{\theta 2}}}$$

As the likelihood function L is an increasing function, for $\alpha^*_{\theta 1}$:

$$\alpha^*_{\theta 1} = \frac{1}{1 + \frac{L(P_2) * \alpha_{\theta 2} + L(P_3) * (1 - \alpha_{\theta 1} - \alpha_{\theta 2})}{L(P_1) * \alpha_{\theta 1}}}$$
$$\leq \frac{1}{1 + \frac{L(y_1) * \alpha_{\theta 2} + L(y_2) * (1 - \alpha_{\theta 1} - \alpha_{\theta 2})}{L(y_1) * \alpha_{\theta 1}}}$$

It reaches the upper bound when

$$P_1 = y_1^-, P_2 = y_1, P_3 = y_2$$

So we got the most conservative prior distribution in the figure 6 (left-hand side) and the corresponding $\alpha^*_{\theta 1}$ as (3.2.3).

Similarly for $\alpha^*_{\theta 1+\theta 2}$:

$$\alpha^*_{\theta 1+\theta 2} = \frac{1}{1 + \frac{L(P_3) * (1 - \alpha_{\theta 1} - \alpha_{\theta 2})}{L(P_1) * \alpha_{\theta 1} + L(P_2) * \alpha_{\theta 2}}} \leq \frac{1}{1 + \frac{L(y_2) * (1 - \alpha_{\theta 1} - \alpha_{\theta 2})}{L(y_1) * \alpha_{\theta 1} + L(y_2) * \alpha_{\theta 2}}}$$

It reaches the upper bound when

$$P_1 = y_1^-, P_2 = y_2^-, P_3 = y_2$$

So we got the most conservative prior distribution in the figure 6 (right-hand side) and the corresponding $\alpha^*_{\theta 1+\theta 2}$ as (3.2.4).

Appendix 3

If we introduce 8 variables with their ranges,

$$\begin{cases} 0 \leq \theta_{PP1} < y, 0 \leq R_1 < r \\ y \leq \theta_{PP2} \leq 1, 0 \leq R_2 < r \\ 0 \leq \theta_{PP3} < y, r \leq R_3 \leq 1 \\ y \leq \theta_{PP4} \leq 1, r \leq R_4 \leq 1 \end{cases}$$

by the mean value theorem for integrals, and similarly as the reasoning in appendix 1 and 2, we would have (as the figure 8, we label the probability masses associated with the four regions as M_1, M_2, M_3, M_4):

$$P(\theta_{PP} < y|\ process\ evidence)$$
$$= \frac{\int_0^1 \int_0^1 (I_{\theta_{PP}<y}(\theta_{PP}) \times [\theta_{PP}+R]^k) g_{<\theta_{PP},R>}(\theta_{PP},R) d\theta_{PP} dR}{\int_0^1 \int_0^1 [\theta_{PP}+R]^k g_{<\theta_{PP},R>}(\theta_{PP},R) d\theta_{PP} dR}$$
$$= \frac{[\theta_{PP1}+R_1]^k M_1 + [\theta_{PP3}+R_3]^k M_3}{[\theta_{PP1}+R_1]^k M_1 + [\theta_{PP3}+R_3]^k M_3 + [\theta_{PP2}+R_2]^k M_2 + [\theta_{PP4}+R_4]^k M_4}$$
$$= \frac{1}{1+\frac{[\theta_{PP2}+R_2]^k M_2 + [\theta_{PP4}+R_4]^k M_4}{[\theta_{PP1}+R_1]^k M_1 + [\theta_{PP3}+R_3]^k M_3}} \leq \frac{1}{1+\frac{y^k M_2 + [\mathrm{y}+r]^k M_4}{[\mathrm{y}+r]^k M_1 + M_3}}$$
$$= \frac{1}{1+\frac{y^k(\gamma_r - M_1) + [\mathrm{y}+r]^k(1-\gamma_\theta - \gamma_r + M_1)}{[\mathrm{y}+r]^k M_1 + (\gamma_\theta - M_1)}}$$

And it is not hard to see the right hand formula is a decreasing function of M_1. And the range of M_1 is $0 \leq M_1 \leq \min(\gamma_\theta, \gamma_p)$, so, when $M_1 = 0$, we reach the upper bound γ_θ^*.

$$P(\theta_{PP} < y|\ process\ evidence) \leq \frac{1}{1+\frac{y^k \gamma_r + [y+r]^k(1-\gamma_\theta - \gamma_r)}{\gamma_\theta}} = \gamma_\theta^*$$

Appendix 4

First let us set the $P(\theta_{PP} < y_1|\ \text{process evidence})$ as our objective function.

We introduce 12 variables with their ranges,

$$\begin{cases} 0 \leq \theta_{PP1} < y_1, 0 \leq R_1 < r \\ y_1 \leq \theta_{PP2} < \mathrm{y}_2, 0 \leq R_2 < r \\ y_2 \leq \theta_{PP5} \leq 1, 0 \leq R_5 < r \\ 0 \leq \theta_{PP3} < y_1, \mathrm{r} \leq R_3 < r_U \\ y_1 \leq \theta_{PP4} < \mathrm{y}_2, \mathrm{r} \leq R_4 < r_U \\ y_2 \leq \theta_{PP6} \leq 1, \mathrm{r} \leq R_6 < r_U \end{cases}$$

By the mean value theorem for integrals and similar reasoning in appendix 1 and 2, we would have (as in figure 10, we label the probability masses associated with the four regions as M_i)

$$P(\theta_{PP} < y_1|\ process\ evidence) = \frac{\int_0^1 \int_0^1 (I_{\theta_{PP}<y1}(\theta_{PP}) \times [\theta_{PP}+R]^k) g_{<\theta_{PP},R>}(\theta_{PP},R) d\theta_{PP} dR}{\int_0^1 \int_0^1 [\theta_{PP}+R]^k g_{<\theta_{PP},R>}(\theta_{PP},R) d\theta_{PP} dR}$$
$$= \frac{[\theta_{PP1}+R_1]^k M_1 + [\theta_{PP3}+R_3]^k M_3}{[\theta_{PP1}+R_1]^k M_1 + [\theta_{PP3}+R_3]^k M_3 + [\theta_{PP2}+R_2]^k M_2 + [\theta_{PP4}+R_4]^k M_4 + [\theta_{PP5}+R_5]^k M_5 + [\theta_{PP6}+R_6]^k M_6}$$
$$\leq \frac{[y_1+r]^k M_1 + [y_1+r_U]^k M_3}{[y_1+r]^k M_1 + [y_1+r_U]^k M_3 + {y_1}^k M_2 + [y_1+r]^k M_4 + {y_2}^k M_5 + [y_2+r]^k M_6}$$
$$= \frac{[y_1+r]^k M_1 + [y_1+r_U]^k (\gamma_{\theta 1} - M_1)}{[y_1+r]^k M_1 + [y_1+r_U]^k (\gamma_{\theta 1} - M_1) + {y_1}^k M_2 + [y_1+r]^k (\gamma_{\theta 2} - M_2) + {y_2}^k (\gamma_r - M_1 - M_2) + [y_2+r]^k (1-\gamma_r - \gamma_{\theta 2} - \gamma_{\theta 1} + M_1 + M_2)}$$

which is a decreasing function in terms of M_1 and M_2 respectively. So, when M_1=M_2=0, we reach the upper bound $\gamma_{\theta 1}^*$:

$$\gamma_{\theta 1}^* = \frac{[y_1+r_U]^k \gamma_{\theta 1}}{[y_1+r_U]^k \gamma_{\theta 1} + [y_1+r]^k \gamma_{\theta 2} + {y_2}^k \gamma_r + [y_2+r]^k (1-\gamma_r - \gamma_{\theta 2} - \gamma_{\theta 1})}$$

The corresponding most conservative prior distribution is shown in figure 11 or 12, depends on the specific values assigned on the parameters.

Note that, when $1 - y_1 \leq r_U < 1$, by the reasoning above, we got the most conservative prior distribution as the figure below.

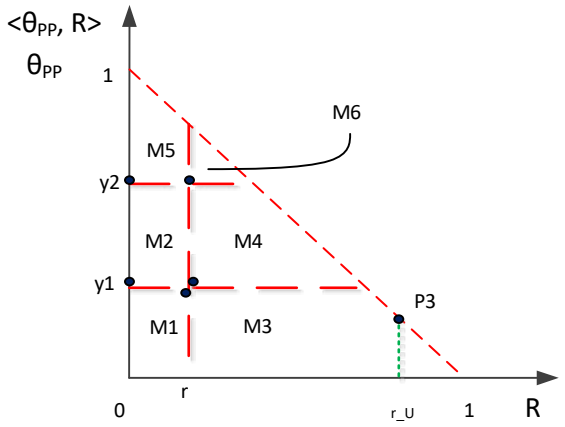

Figure A.1 a most conservative prior dist. giving the counter-intuitive result

In this case, the P_3 will always win out. As the P_3 point is below y_1, we will have $P(\theta_{PP} < y_1 |$ process evidence$) > P(\theta_{PP} < y_1)$, which is the "counter-intuitive" result again.

Similarly as the reasoning above, it is not hard to draw the most conservative prior distribution for the objective function $P(\theta_{PP} < y_2 |$ process evidence$)$ as below:

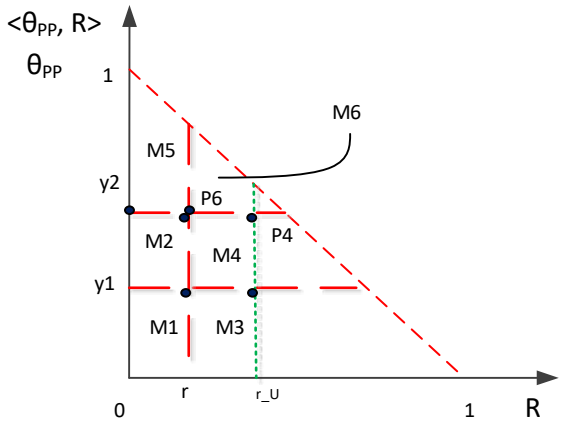

Figure A.2 a most conservative prior dist. giving the counter-intuitive result

When seeing good process evidence, the P_4 point will always win out of the other points on the joint distribution, no matter how the r_u varies in its range. In other words, the mass at other points will move to the P_4 point. As the P_4 point is below y_2, we will have $P(\theta_{PP} < y_2 | \, process \; evidence) > P(\theta_{PP} < y_2)$, which is the "counter-intuitive" result again.

References

Aven, T. and Zio, E. (2011). Some considerations on the treatment of uncertainties in risk assessment for practical decision making. Reliability Engineering & System Safety, 96(1), 64-74.

Bertolino, A. and L. Strigini (1998). "Assessing the risk due to software faults: estimates of failure rate vs evidence of perfection." Journal of Software Testing, Verification and Reliability **8**(3): 155-166.

Bishop, P., R. Bloomfield, B. Littlewood, A. Povyakalo and D. Wright (2011). "Towards a formalism for conservative claims about the dependability of software-based systems." IEEE Trans Software Engineering **37**(5): 708-717.

Boeing (2015). Statistical Summary of Commercial Airplane Accidents, Worldwide Operations, 1959-2014. Seattle, Aviation Safety, Boeing Commercial Airplanes.

Bunea, C., T. Charitos, R. M. Cooke and G. Becker (2005). "Two-stage Bayesian models - application to ZEDB project." Reliability Engineering and System Safety **90**(2): 123-130.

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

Cooke, R., C. Bunea, T. Charitos and T. A. Mazzuchi (2002). Mathematical review of ZEDB two-stage Bayesian models. Delft, Department of Mathematics, Delft University of Technology. **Report 02-45**.

Eckhardt, D. E., A. K. Caglayan, J. C. Knight, L. D. Lee, D. F. McAllister, M. A. Vouk and J. P. J. Kelly (1991). "An experimental evaluation of software redundancy as a strategy for improving reliability." IEEE Trans Software Eng **17**(7): 692-702.

Eckhardt, D. E. and L. D. Lee (1985). "A Theoretical Basis of Multiversion Software Subject to Coincident Errors." IEEE Trans. on Software Engineering **11**: 1511-1517.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin (2013). Bayesian Data Analysis, CRC Press.

HSE (2013). GDA Step 4 and Close-out for Control and Instrumentation Assessment of the EDF and AREVA UK EPR Reactor. Bootle, Health and Safety Executive, Office for Nuclear Regulstion.

Kaplan, S. (1983). "On a two-stage Bayesian procedure for determining failure rates from experimental data." IEEE Trans on Power Apparatus and Systems **1**(195-202).

Knight, J. C. and N. G. Leveson (1986). "Experimental evaluation of the assumption of independence in multiversion software." IEEE Trans Software Engineering **12**(1): 96-109.

Lehmann, E. L. and G. Cassella (2003). Theory of Point Estimation, Springer.

Littlewood, B. and D. R. Miller (1989). "Conceptual Modelling of Coincident Failures in Multi-Version Software." IEEE Trans on Software Engineering **15**(12): 1596-1614.

Littlewood, B., P. Popov and L. Strigini (2001). "Modelling software design diversity - a review." ACM Computing Surveys **33**(2): 177-208.

Littlewood, B. and A. Povyakalo (2013). "Conservative Reasoning about the Probability of Failure on Demand of a 1-out-of-2 Software-Based System in Which One Channel Is 'Possibly Perfect'." IEEE Trans Software Engineering **39**(11): 1521-1530.

Littlewood, B. and J. Rushby (2012). "Reasoning about the reliability of diverse two-channel systems in which one channel is 'possibly perfect'." IEEE Trans Software Engineering **38**(5): 1178-1194.

Littlewood, B. and L. Strigini (1993). "Validation of ultra-high dependability for software-based systems." CACM **36**(11): 69-80.

Littlewood, B. and D. Wright (1997). "Some conservative stopping rules for the operational testing of safety-critical software." IEEE Trans Software Engineering **23**(11): 673-683.

Littlewood, B. and D. Wright (2007). "The use of multi-legged arguments to increase confidence in safety claims for software-based systems: a study based on a BBN of an idealised example." IEEE Trans Software Engineering **33**(5): 347-365.

NRC (2003). Handbook of parameter estimation for probabilistic risk assessment. US Nuclear Regulatory Commission, Washington, DC, Report No. NUREG/CR-6823.

NRC (2017). Guidance on the Treatment of Uncertainties Associated with PRAs in Risk-informed Decision Making: Main Report. US Nuclear Regulatory Commission, Washington, DC, Report No. NUREG-1855 Revision-1.

PowerTech, V. (2010). Centralised reliability and events database (ZEDB) - reliability data for nuclear power plant components. Essen, VGB PowerTech. **VGB-TW805e-11**.

RTCA (1992). Software considerations in airborne systems and equipment certification, DO-178B, Requirements and Technical Concepts for Aeronautics.

Rushby, J. (2009). Software verification and system assurance. 7th IEEE International Conference on Software Engineering and Formal Methods (SEFM), Hanoi, Vietnam, IEEE Computer Society.

Strigini, L. and A. Povyakalo (2013). Software fault-freeness and reliability predictions. SAFECOMP 2013, 32nd International conference on Computer Safety, Reliability and Security. Toulouse.

Vaurio, J. K. and K. E. Jänkälä (2006). "Evaluation and comparison of estimation methods for failure rates and probabilities." Reliability Engineering and System Safety **91**(2): 209-221.

Wood, R. T. and R. Belles (2010). Diversity Strategies for Nuclear Power Plant Instrumentation and Control Systems. Washington, DC, US Nuclear Regulatory Commission, NUREG/CR-7007.

Zhao, X., Littlewood, B, Povyakalo, A. A., Wright, D., Strigini, L. (2017). "Modeling the probability of failure on demand (pfd) of a 1-out-of-2 system in which one channel is “quasi-perfect”." Reliability Engineering and System Safety **158**: 230-245.

Zhao, X., Littlewood, B., Povyakalo, A. A. & Wright, D. (2015). Conservative Claims about the Probability of Perfection of Software-based Systems. The 26th IEEE International Symposium on Software Reliability Engineering (ISSRE2015). Gaithersburg, MD, USA, IEEE Computer Society**:** 130-140.

Zio, E. and N. Pedroni (2013). Literature review of methods for representing uncertainty. Toulouse, Cahiers de la Sécurité Industrielle, Foundation for an Industrial Safety Culture. **2013-03**.

Acknowledgements

We are grateful to our anonymous reviewers for their careful readings of our paper. Their extensive criticisms and many helpful suggestions have improved it greatly.

The work reported here was supported in part by the UK C&I Nuclear Industry Forum (CINIF). The views expressed in this paper are those of the author(s) and do not necessarily represent the views of the members of the C&I Nuclear Industry Forum (CINIF). CINIF does not accept liability for any damage or loss incurred as a result of the information contained in this paper.