Modelling the probability of failure on demand (*pfd*) of a 1-out-of-2 system in which one channel is "quasi-perfect"

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

Centre for Software Reliability, City University London

Abstract

Our earlier work proposed ways of overcoming some of the difficulties of lack of independence in reliability modeling of 1-out-of-2 software-based systems. Firstly, it is well known that aleatory independence between the failures of two channels *A* and *B* cannot be assumed, so system *pfd* is not a simple product of channel *pfd*s. However, it has been shown that the probability of system failure *can* be bounded conservatively by a simple product of pfd_A and pnp_B (probability not perfect) in those special cases where channel *B* is sufficiently simple to be possibly perfect. Whilst this "solves" the problem of aleatory dependence, the issue of epistemic dependence remains: An assessor's beliefs about unknown pfd_A and pnp_B will not have them independent. Recent work has partially overcome this problem by requiring only *marginal* beliefs – at the price of further conservatism. Here we generalize these results. Instead of "perfection" we introduce the notion of "quasi-perfection": a small *pfd* practically equivalent to perfection (e.g. yielding very small chance of failure in the entire life of a fleet of systems). We present a conservative argument supporting claims about system *pfd*. We propose further work, e.g. to conduct "what if?" calculations to understand exactly *how* conservative our approach might be in practice, and suggest further simplifications.

KEY WORDS: *Fault-free software; program perfection; quasi-perfection; probability of perfection; 1-out-of-2 system reliability; software diversity*

1 Introduction and background

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. For example, for flight-critical avionics systems in civil transport airplanes there is a requirement of less than 10^{-9} probability of failure per hour of operation (FAA 1988). Some demand-based systems have similarly stringent requirements: e.g. the claimed probability of failure on demand (*pfd*) for the combined control and instrumentation safety systems on the UK European Pressurised Reactor (UK EPR) is 10^{-9} (HSE 2011). To achieve this kind of ultra-high dependability is not only a difficult task of design and implementation, but poses even harder problems of assessment. Direct black-box operational testing, for example, would require infeasible times on testing (Littlewood and Strigini 1993).

Design diversity has been proposed as a promising way of achieving high dependability for software-based systems. The intuitive explanation 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 2013) with apparently no critical failure (note, however, that these continuously operating systems have a different architecture from the 1-o-o-2 *on-demand*[1] 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).

On the other hand, evidence like this is only available after the fact. Assessing the reliability of such a design-diverse system *before* it is deployed remains a very difficult problem. It is difficult because it requires an understanding of (and a formal representation of) two different kinds of uncertainty: *aleatory* uncertainty and *epistemic* uncertainty. Informally, the first of these concerns uncertainty *in the world* – in our case, uncertainty about the channel failures and their impact upon system failure. Because it is a property of "the world out there", this kind of uncertainty can be thought of as irreducible. In contrast, epistemic uncertainty can be thought of as uncertainty *about the world* – for example, about the values of parameters in our model of aleatory uncertainty in the world. It may, for example, involve statistical inference; it may thus be reducible (by acquiring more evidence), but generally cannot be eliminated completely.

We begin by looking first at the problem of aleatory uncertainty concerning dependence of channel failures in our problem. We know, from experimental work (Knight and Leveson 1986, Eckhardt, Caglayan et al. 1991) and theoretical modeling (Eckhardt and Lee 1985, Littlewood and Miller 1989) that we cannot claim in general that there is independence between the failures of multiple software-based channels of a system. Thus for a 1-o-o-2 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 from extensive testing, say P_A and P_B, we cannot simply multiply them and claim the system *pfd* is $P_A \times P_B$.

In recent work by Littlewood and Rushby (Littlewood and Rushby 2012), hereafter LR, the authors proposed a new way to reason about the reliability of a special kind of 1-o-o-2 systems architecture. Here channel *A* is conventionally engineered and

[1] The requirements for critical flight control systems are expressed in terms of a continuous-time failure rate – i.e. 10^{-9} per hour of operation (FAA 1988). Interestingly, though, accident statistics are reported in (Boeing 2013) in terms of numbers of departures, which equate to numbers of flights. It could be argued that a probability of failure per demand, ie. per flight, is the most appropriate way of expressing a reliability requirement here..

presumed to be sufficiently complex that it will contain design faults. It must therefore be assumed that it will (eventually) exhibit failures, so its reliability will be expressed as a claim about its probability of failure on demand (say pfd_A). Channel B on the other hand has been designed to be extremely simple and has been extensively analyzed. It is therefore open to a claim of being "possibly perfect"; the claim about this channel is consequently a probability of non-perfection, pnp_B. Littlewood and Rushby show that:

$$P(\text{system fails on randomly selected demand} \mid pfd_A, pnp_B) \leq pfd_A \times pnp_B \qquad (1)$$

The result depends on the fact that there is conditional independence between the events "A fails on a randomly selected demand" and "B is not perfect," given that the probabilities of these events, respectively pfd_A and pnp_B, are known. This is a conservative bound for the system's probability of failure on a randomly selected demand (pfd_{sys}), and the conservatism arises by assuming 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 very small (bound on) pfd_{sys} (cf the product of pfd_A and pfd_B above which has the same intent, but requires the generally unsupportable assumption of independence of channel failures). The expectation is that stronger claims about the *pfd* of a system can be made than would be possible via direct, black-box evaluation.

The LR result, (1), can be seen as "solving" our problem of aleatory uncertainty here – concerning in-the-world dependence between failures – but it is, of course, a bound on a *conditional* probability. In reality, an assessor would not know pfd_A and pnp_B with certainty. This brings us to the problem of *epistemic uncertainty* about the numerical values of these model parameters.

In principle, an assessor could represent his epistemic uncertainty (i.e. his subjective beliefs) here via a complete bivariate distribution for the two unknowns. In practice people find this kind of thing very difficult, if not impossible. A major source of this difficulty concerns, as in the case of aleatory uncertainty above, dependence. Even if assessors are able to make informed statements about their *marginal* beliefs about the two parameters, individually, they will usually be unable to say anything about their *dependence.*

Littlewood and Povyakalo address this problem in (Littlewood and Povyakalo 2013), hereafter LP. They obtain results 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 his beliefs about the two numbers. The price paid here is further conservatism, in addition to that arising from the LR result.

The results of LR and LP, then, have reduced the problem of assessing the *pfd* of this kind of special 1oo2 system to one concerning simply marginal beliefs about the parameters pfd_A and pnp_B. There is a large literature on the assessment of *pfd* from statistical analysis of operational tests, 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. That leaves pnp_B, which was the subject of our earlier paper (Zhao 2015).

The question upon which we concentrated in that paper was what can be claimed about probability of perfection from seeing many failure-free tests. We develop a probability model for this problem, and illustrate it with some numerical examples. Our approach starts from the premise that real assessors can generally only provide *limited prior belief*, rather than a complete distribution; for example, a probability mass at the origin (representing prior confidence in perfection, which is also required in later sections of this paper), and one percentile for the rest of the distribution. In the face of this difficulty, our approach is *conservative*: from the many (generally an infinite number of) distributions that satisfy the limited prior constraints, we choose the one(s) that give the most conservative results for the system's posterior *pfd*. Unfortunately, such results are often *very* conservative – in fact too conservative to be useful. In fact, in the worst case, confidence in perfection does not increase even after observing an *infinite* number of successful tests. We comment that this is because, whilst extensive failure-free working may be a result of a program's perfection, it could also be because the program – although not perfect – has a very small *pfd*. We propose some ways around this problem, essentially by pruning the large class of allowable prior distributions by excluding ones that seem "unreasonable" in general ways. Of course, in a real application, the assessor would need to accept the reasonableness of these further constraints on his prior beliefs.

In the work to be described in the remainder of this paper we propose a different way around this difficulty. The idea is to *exploit* the fact that "perfection" and "extremely small *pfd*" are effectively indistinguishable as explanations for extensive failure-free working. We introduce the notion of "quasi-perfection" of a channel: that the *pfd* of the channel is smaller than some small given number ε. For example, ε could be chosen so that over the entire lifetime of the system (or fleet of systems) there would be only a small chance of failure, i.e. the lifetime behavior of the system would be anticipated to be identical to that of a perfect one.[2]

In this paper we show that for some combinations of parameters, the conservative bounds for reliability of a 1-o-o-2 system, based on claims about quasi-perfection of one of its channel, are more sensitive to Bayesian update of their parameters and *less conservative* than the bounds based on claims about pure perfection in our earlier work.

The idea in this work is a generalization of an observation by Strigini and Povyakalo (Strigini and Povyakalo 2013). The usefulness of probability of perfection transcends its application to the LR model of a 1-out-of-2 system: in fact it is a lower bound on the probability of failure-free operation of a system over *any arbitrarily long period of operation* (Bertolino and Strigini 1998). The reliability of a program, $R(t)$ – its probability of surviving failure-free for time t – satisfies $R(t) \geq P$(program is perfect) *however large* t *is*. Strigini and Povyakalo observed that forms of mathematical arguments based on probability of perfection can be extended to using probabilities of "quasi-perfection", a notion that we exploit in what follows.

2 A new bound for the reliability of a 1oo2 system based on the possible "quasi-perfection" of one channel

Our interest centres on the probability of failure of a 1-out-of-2 system with channels *A* and *B*. Our knowledge of the channels is such that we shall make a claim about *probability of failure on demand* for channel *A* (pfd_A), and *probability of not-quasi-perfect* for channel *B* ($pnqp_B$).

[2] Consider the example of a single channel of a nuclear reactor protection system. We might anticipate something like 2 demands on the protection system on average per year, with an anticipated lifetime of 50 years. For 99% confidence of seeing no failures in the expected 100 demands, ε should be about 10^{-4}.

It is useful to start here with an account of the underlying probability model. We shall do this via that cliché of elementary probability theory – the random selection of coloured balls from urns.

Consider first urn *A*. This is filled with balls representing all possible *demands* that the system can encounter. Balls are either white or black. A black ball represents a demand that channel *A* cannot execute correctly, i.e. *A* fails on such a demand. White balls represent demands on which *A* succeeds. Selection of balls – i.e. demands – is determined by the operational profile, and successive selections are statistically independent. In general, different balls have different probabilities of selection. For a ball randomly selected via the operational profile, the probability that it is black is simply the probability of failure on demand of channel *A*, pfd_A. A “frequentist” interpretation of this probability is that it is the limiting relative frequency of demands for which channel *A* fails in an infinite sequence of independently selected demands Note that in this model some care needs to be taken in the definition of “demand” to satisfy the requirement that successive ball-drawings are independent. It is a property of the world rather than of the computer system. In the example of the reactor protection system, a demand will be a *trajectory* through the input space (of temperatures, pressures, flow-rates, etc), from departure from safe operation to shut-down. For example, a demand will not simply be a control loop cycle. Demands will be separated by long periods of safe operation, so it is reasonable to expect that a demand now will be independent of one several months ago.

Now consider the second urn, *B*. This is filled with balls representing all possible *programs* that could be developed to solve the problem at hand, using the development process that characterizes *B*. Each such program will be either quasi-perfect (*qp*) or not quasi-perfect (*nqp*). Let white balls represent *qp* programs, black balls *nqp* programs. The development of a program can now be seen as the random selection of a program from this urn. Again, different balls – programs – will have different probabilities of selection. For a ball selected at random – i.e. a program developed using the process *B* – the probability that it is black is simply the probability that the program is not quasi-perfect, $pnqp_B$. Once again, a frequentist interpretation of this probability is that it is the limiting relative frequency of *nqp* programs in an infinite sequence of independently selected (i.e. developed) programs.

These interpretations of the parameters, pfd_A and $pnqp_{\boldsymbol{B}}$, of course involve thought experiments. We can never actually *see* an infinite number of demands, much less an infinite number of programs. Readers may be more comfortable with the limiting relative frequency interpretation of pfd_A than that of $pnqp_B$. The latter is essentially the same idea as that used in LR (Littlewood and Rushby 2012) to define pnp_B (the probability that channel *B* is not perfect). The first uses of the notion of “randomly selected” program seem to be due to Eckhardt and Lee (Eckhardt and Lee 1985) and, in a later generalization, Littlewood and Miller (Littlewood and Miller 1989). In the old experiments on multi-version software – see e.g. (Knight and Leveson 1986, Eckhardt, Caglayan et al. 1991) – the multiple software versions were modeled as random samples from populations of programs.

For our purposes here, the kind of hypothetical replication of programs, to form the population from selection would take place in our thought experiment, would require that they all aim to solve *this particular problem* under examination, using *development teams* of comparable competence and experience, and using the same *development process* (i.e. the same software engineering practices). This allows us to interpret the limiting relative frequency of *qnp* programs in a sequence of selected programs to be the *pqnp* for *this problem*, and *a similar development team*, using the *same development processes*.

The experiment we conduct now is the *independent* selection of a ball from urn *A* and a ball from urn *B*. The events "ball from *A* is black" and "ball from *B* is black" are then assumed *conditionally* independent for given values of pfd_A and $pnqp_B$ (say $pfd_A=p_A$, $pnqp_B=p_B$, respectively).

In the terminology of our reliability problem: the events "channel *A* fails on a randomly selected demand" and "program *B* is not quasi-perfect" are conditionally independent given $pfd_A=p_A$, $pnqp_B=p_B$, say. This conditional independence assumption is the key to the proof of the theorem that follows. It is similar to the conditional independence assumption in LR; in that paper it concerns independence between "*A* fails" and "*B* not perfect".

Such conditional independence is, we believe, a realistic assumption for this situation. Imagine that you have built program *B* (i.e., in our thought experiment, selected it randomly from the population of process-*B* programs). You now execute a randomly selected demand on program *A*. You should not expect the outcome of the demand selection to be influenced by the outcome of the *B* selection. It follows that whether or not *A* fails should not be influenced by whether or not *B* was quasi-perfect.

Note that the events "channel *A* fails on a randomly selected demand" and "program *B* is not quasi-perfect" are *not unconditionally independent* in general. Informally, seeing *A* fail on a demand may suggest to an assessor that the problem being solved by both programs is a "difficult" one, and this may change his subjective belief that *B* will be quasi-perfect. That is, learning something about *A*'s probability of failure on demand (e.g. by seeing an *A* failure), may tell us something about *B*'s probability of being not quasi-perfect. This is an issue of *epistemic* dependence between the assessor's beliefs about the model parameters, pfd_A and $pnqp_B$, which we shall address later in the paper.

We can now return to the object of our interest: the probability of *system* failure on a randomly selected demand. We begin with the conditional probability of failure given $pfd_A=p_A$, $pnqp_B=p_B$:

Theorem 1

$$P(\text{System fails} \mid pfd_A = p_A, pnqp_B = p_B) \le \varepsilon.(1-p_B) + p_A.p_B \qquad (2)$$

Proof:

$$\begin{aligned}
&P(\text{System fails} \mid pfd_A = p_A, pnqp_B = p_B)\\
&= P(\text{System fails} \mid A \text{ fails, } B \text{ is } nqp, pfd_A = p_A, pnqp_B = p_B)\times\\
&P(A \text{ fails, } B \text{ is } nqp \mid pfd_A = p_A, pnqp_B = p_B)\\
&+P(\text{System fails} \mid A \text{ fails, } B \text{ is } qp, pfd_A = p_A, pnqp_B = p_B)\times\\
&P(A \text{ fails, } B \text{ is } qp \mid pfd_A = p_A, pnqp_B = p_B)\\
&+P(\text{System fails} \mid A \text{ succeeds, } B \text{ is } nqp, pfd_A = p_A, pnqp_B = p_B)\times\\
&P(A \text{ succeeds, } B \text{ is } nqp \mid pfd_A = p_A, pnqp_B = p_B)\\
&+P(\text{System fails} \mid A \text{ succeeds, } B \text{ is } qp, pfd_A = p_A, pnqp_B = p_B)\times\\
&P(A \text{ succeeds, } B \text{ is } qp \mid pfd_A = p_A, pnqp_B = p_B)
\end{aligned} \qquad (3)$$

and the last two terms on the right hand side of the expansion (3) are zero trivially, since if A succeeds the 1-out-of-2 system cannot fail.

Now, if B is not *qp*, it is conservative to assume that it fails whenever A does, so the first term on the right hand side of the expansion (3) is

$$\leq 1\times P(A \text{ fails}, B \text{ is } nqp \mid pfd_A = p_A, pnqp_B = p_B)$$
$$= P(A \text{ fails} \mid pfd_A = p_A, pnqp_B = p_B)\times P(B \text{ is } nqp \mid pfd_A = p_A, pnqp_B = p_B)$$
$$= p_A \times p_B \tag{4}$$

because "*A* fails" and "*B* is *nqp*" are independent given $pfd_A = p_A$ and $pnqp_B = p_B$.

The second term in the expansion is

$$P(\text{System fails} \mid A \text{ fails}, B \text{ is } qp, pfd_A = p_A, pnqp_B = p_B)\times$$
$$P(A \text{ fails}, B \text{ is } qp \mid pfd_A = p_A, pnqp_B = p_B)$$
$$= P(A \text{ and } B \text{ fail} \mid A \text{ fails}, B \text{ is } qp, pfd_A = p_A, pnqp_B = p_B)\times p_A \times (1-p_B)$$

(where we have relabelled the event "System fails" as "A and B fail", without change of meaning) because "*A* fails" and "*B* is *nqp*" are independent given $pfd_A = p_A$ and $pnqp_B = p_B$. Now making explicit the conditioning on event "A fails" in this expression:

$$= \frac{P(A \text{ and } B \text{ fail} \mid B \text{ is } qp, pfd_A = p_A, pnqp_B = p_B)\times p_A \times (1-p_B)}{P(A \text{ fails} \mid B \text{ is } qp, pfd_A = p_A, pnqp_B = p_B)}$$

and considering in the numerator that "A and B fail" is a subset of "B fails":

$$\leq \frac{P(B \text{ fails} \mid B \text{ is } qp, pfd_A = p_A, pnqp_B = p_B)\times p_A \times (1-p_B)}{P(A \text{ fails} \mid B \text{ is } qp, pfd_A = p_A, pnqp_B = p_B)}$$
$$= \frac{P(B \text{ fails} \mid B \text{ is } qp, pfd_A = p_A, pnqp_B = p_B)\times p_A \times (1-p_B)}{p_A}$$

because "*A* fails" and "*B* is *qp*" are independent given $pfd_A = p_A, pnqp_B = p_B$; and seeing that if B is *qp*, its probability of failure is ≤ε for any (p_A,p_B):

$$\leq \frac{\varepsilon.p_A.(1-p_B)}{p_A}$$
$$= \varepsilon.(1-p_B) \tag{5}$$

So finally we have, by substituting (4) and (5) into (3):

$$P(\text{System fails} \mid pfd_A = p_A, pnqp_B = p_B) \leq \varepsilon.(1-p_B) + p_A.p_B$$

QED

Note that this result is a generalization of the one in LR (Littlewood and Rushby 2012): we obtain this earlier result by putting ε=0.

3 Conservative reasoning about the epistemic uncertainty

The result above concerns what happens at the aleatory level in the model. In practice, of course, the parameters of the model will not be known with certainty. Ideally, an assessor would describe his *epistemic uncertainty* about these unknowns – pfd_A and $pnqp_B$ – in terms of a complete bivariate distribution:

$$F_{pfd_A,pnqp_B}(p_A,p_B)=P(pfd_A\le p_A,pnqp_B\le p_B) \tag{6}$$

The *unconditional* probability of system failure is then

$$\begin{aligned}&P(\text{System fails})=E_{pfd_A,pnqp_B}\left(P(\text{System fails}\,|\,pfd_A=p_A,pnqp_B=p_B)\right)\\&\le E_{pfd_A,pnqp_B}\left(\varepsilon.(1-pnqp_B)+pfp_A.pnqp_B\right)\\&=\int\int\left(\varepsilon.(1-p_B)+p_A.p_B\right)dF_{pfd_A,pnqp_B}(p_A,p_B)\end{aligned} \tag{7}$$

In reality, it is unlikely that a real-world assessor would be willing or able to offer such a complete bivariate distribution to represent his beliefs about the unknowns of the model. In particular, it is known that people find it hard to express the *dependence* between their beliefs. In this section, therefore, we obtain some results that are based only on *marginal* beliefs; the price paid for this simplification of the assessor's task is further conservatism in the results. These results about quasi-perfection generalise those of LP (Littlewood and Povyakalo 2013) about perfection.

Once again we expect that an assessor will not be able to provide *complete* subjective distributions, even to represent his marginal beliefs about the unknown parameters. As in LP, then, the different theorems here give results for the system *pfd* based on different kinds of *limited* marginal beliefs that the assessor may be able to express about the unknown parameters.

3.1 Conservative bounds on mean system pfd

Theorem 2

Assume the assessor could only give us a single percentile of his marginal belief for each distribution:

$$P(pfd_A\le P_A)=1-\alpha_A \tag{8}$$

$$P(pnqp_B\le P_B)=1-\alpha_B \tag{9}$$

where $0\le\alpha_A,\alpha_B\le 1$, and we assume $P_A\ge\varepsilon$, i.e that the value of the assessor's 100(1-α_A)-percentile for *pfd* of the *A* channel exceeds[3] the threshold ε used to define quasi-perfection of the *B* channel.

Then

$$E(pfd_{sys})\le\varepsilon+(P_A-\varepsilon)P_B+(1-P_A)P_B\alpha_A+(P_A-\varepsilon)(1-P_B)\alpha_B+(1-P_A)(1-P_B)\alpha_m \tag{10}$$

where $\alpha_m=\min(\alpha_A,\alpha_B)$. This bound equates either to

$$E(pfd_{sys})\le\varepsilon+(P_A-\varepsilon)P_B+(1-P_A)\alpha_A+(P_A-\varepsilon)(1-P_B)\alpha_B \tag{11}$$

if $\alpha_A\le\alpha_B$, or to

$$E(pfd_{sys})\le\varepsilon+(P_A-\varepsilon)P_B+(1-P_A)P_B\alpha_A+(1-\varepsilon)(1-P_B)\alpha_B \tag{11a}$$

if $\alpha_B\le\alpha_A$

(see Appendix for proof).

[3] A modified but similar bound, which we omit here for brevity, applies in the other, we think less likely, case.

Example 1

The assessor has chosen ε=10^{-7} to define quasi-perfection, i.e. if pfd_B is smaller than this, he will regard channel B to be quasi-perfect. If the assessor is 95% confident that pfd_A is smaller than 10^{-5}, and 95% confident that $pnqp_B$ is smaller than 10^{-2}, we have from (11):

$$E(pfd_{sys}) \leq \varepsilon + (10^{-5} - 10^{-7})10^{-2} + (1 - 10^{-5})0.05$$

$$+(10^{-5} - 10^{-7})(1 - 10^{-2})0.05 \approx 0.05 \tag{12}$$

which is, of course, *very* conservative. It is easy to see that the value of the mean system pfd in (11) is dominated by the smallest doubt, i.e. α_A.

To overcome this problem, we consider in Theorem 3 a case where the assessor is able to provide, in addition to the percentile constraints above, upper bounds for the parameters about which he is *certain*:

Theorem 3

If, in addition to the beliefs (8), (9) in Theorem 2, and retaining assumption $P_A \geq \varepsilon$, the assessor also believes:

$$P(pfd_A < P_A^U) = 1 \tag{13}$$

$$P(pnqp_B < P_B^U) = 1 \tag{14}$$

That is, he has 1-α_A confidence that the *pfd* of channel *A* is smaller than P_A, but he is certain that it is smaller than $P^U{}_A$; and he has 1-α_B confidence that the *pnqp* of channel *B* is smaller than P_B, but he is certain that it is smaller than $P^U{}_B$, then we have

$$E(pfd_{sys}) \leq \varepsilon + (P_A - \varepsilon)P_B + (P_A^U - P_A)P_B\alpha_A + (P_A - \varepsilon)(P_B^U - P_B)\alpha_B + (P_A^U - P_A)(P_B^U - P_B)\alpha_m \tag{15}$$

where $\alpha_m = \min(\alpha_A, \alpha_B)$. This bound equates either to

$$E(pfd_{sys}) \leq \varepsilon + (P_A - \varepsilon)P_B + (P_A^U - P_A)P_B^U\alpha_A + (P_A - \varepsilon)(P_B^U - P_B)\alpha_B \tag{15a}$$

if $\alpha_A \leq \alpha_B$, or to

$$E(pfd_{sys}) \leq \varepsilon + (P_A - \varepsilon)P_B + (P_A^U - P_A)P_B\alpha_A + (P_A^U - \varepsilon)(P_B^U - P_B)\alpha_B$$

(15b) if $\alpha_B \leq \alpha_A$

(see Appendix for proof).

Example 2

This adds the two pieces of "certainty" belief to Example 1. That is, as before, ε=10^{-7}, the assessor is 95% confident that pfd_A is smaller than 10^{-5}, and 95% confident that $pnqp_B$ is smaller than 10^{-2}. Additionally, he is certain that pfd_A is no worse than 10^{-3}, and $pnqp_B$ is no worse than 10^{-1}. Then we have, by substitution into (15):

$$E(pfd_{sys}) \leq 0.00000519355 \tag{16}$$

Clearly this bound, using the two further constraints, is much better than that in Example 1, (12). In fact it is more than an order of magnitude better than the product of the assessor's worst case values for *pfd*$_A$ and *pnqp*$_B$ (respectively, 10^{-3} and 10^{-1}): which is the LR result (i.e. ε=0) for this worst case situation.

Of course, the numbers chosen for these examples are merely illustrative, and not meant to represent what actual assessors would believe in real-life situations. However, the reader may think that the choice of the "certainty" bounds here is itself conservative: if an assessor is 95% confident that *pfd*$_A$ is smaller than 10^{-5}, it may be reasonable for him to believe it is certainly not two orders of magnitude worse than that (and similarly for *pnqp*$_B$ not more than one order of magnitude worse).

In the previous theorems, the assessor expressed his beliefs about the parameters in terms of (a small number of) *percentiles* of his marginal distributions. The following theorem treats the case where the assessor's limited beliefs are expressed in terms of the first two moments (mean and variance) of these distributions.

Theorem 4

$$E(pfd_{sys}) \leq E[\varepsilon \times (1 - pnqp_B) + pfd_A \times pnqp_B]$$

$$= \varepsilon - \varepsilon \times E(pnqp_B) + E(pfd_A \times pnqp_B)$$

$$< \varepsilon - \varepsilon \times E(pnqp_B) + \sqrt{(E(pfd_A)^2 + Var(pfd_A))(E(pnqp_B)^2 + Var(pnqp_B))} \tag{17}$$

$$< \varepsilon - \varepsilon \times E(pnqp_B) + (E(pfd_A)\ + SD(pfd_A))(E(pnqp_B)\ + SD(pnqp_B)) \tag{18}$$

(see Appendix for proof).

Example 3

If we know,

$$SD(pfd_A) < 4E(pfd_A)\ and\ SD(pnqp_B) < 4E(pnqp_B)$$

Then we have from (17)

$$E(Pfd_{sys}) < \varepsilon - \varepsilon \times E(pnqp_B) + 17 \times E(pfd_A)\ E(pnqp_B)$$

Finally, for this subsection, the next theorem treats the case where the assessor provides a single percentile, and a special bound on the mean, for each marginal distribution.

Theorem 5

If

$$P(pfd_A \leq P_A) = 1 - \alpha_A$$

$$P(pnqp_B \leq P_B) = 1 - \alpha_B$$

and

$$E(pfd_A) \leq P_A \leq \alpha_A$$

$$E(pnqp_B) \leq P_B \leq \alpha_B$$

Then

$$E(pfd_{sys}) \leq \varepsilon - \varepsilon \times E(pnqp_B) +$$
$$\sqrt{(1+P_A)E(pfd_A) - \alpha_A P_A} \times$$
$$\sqrt{(1+P_B)E(pqnp_B) - \alpha_B P_B} \tag{19}$$
$$\leq \varepsilon - \varepsilon \times E(pnqp_B) + \sqrt{P_A^2 + P_A(1-\alpha_A)} \times \sqrt{P_B^2 + P_B(1-\alpha_B)} \tag{20}$$

(see Appendix for proof).

Example 4

If, as in example 1:

$P_A = 10^{-5}, \alpha_A = 0.05$ and $P_B = 10^{-2}, \alpha_B = 0.05, \varepsilon = 10^{-7}$

and the assessor also told us that $E(pfd_A) \leq P_A, E(pqnp_B) \leq P_B$, then we have

$$E(Pfd_{sys}) \leq 10^{-7} - 10^{-7} \times 10^{-2} +$$
$$\sqrt{10^{-10} + 10^{-5} \times 0.95} \times \sqrt{10^{-4} + 10^{-2} \times 0.95} = 0.000302094 \tag{21}$$

which is again better than the result in example 1 and worse than the one in example 2.

Of course, both Theorem 3 and Theorem 5 need extra information compared with Theorem 2. However, it could be argued that the extra information may be easy to justify: there is no need to know the exact mean values in the theorem, merely that the marginal means are smaller than the corresponding percentiles (even if the exact values of some of these are not known to the assessor).

3.2 Confidence bounds for system pfd

Instead of obtaining a mean value, we could also – as in LP – get a conservative confidence bound for system pfd using only an assessor's marginal knowledge of *pfd_A* and *pnqp_B*. The following theorem does this using only a single percentile for each distribution.

Theorem 6

Given a single percentile of the marginal belief for each distribution,

$$P(pfd_A \leq P_A) = 1 - \alpha_A \tag{22}$$
$$P(pnqp_B \leq P_B) = 1 - \alpha_B \tag{23}$$

we have

$$P\left(pfd_{sys} < \varepsilon \times (1-P_B) + P_A \times P_B\right) > 1 - (\alpha_A + \alpha_B) \tag{24}$$

Example 5

If, as before, $P_A = 10^{-5}, \alpha_A = 0.05$ and $P_B = 10^{-2}, \alpha_B = 0.05, \varepsilon = 10^{-7}$

then we have

$$P\left(pfd_{sys} < 1.99 \times 10^{-7}\right) \geq 0.9 \tag{25}$$

(see Appendix for proof)

Example 6

If the assessor can provide two (or more) percentiles for each distribution, then multiple conservative percentiles can be generated for the distribution of *pfd_{sys}*. So if, in addition to the two percentiles above, the assessor is 99% confident that *pfd_A* is smaller than 10^{-3}, and 99.9% confident that *$pnqp_B$* is smaller than 10^{-1}, the following conservative percentiles apply to his beliefs about the system pfd:

1. $P(pfd_{sys} < 1.0009 \times 10^{-4}) \geq 0.989$
2. $P(pfd_{sys} < 1.09 \times 10^{-6}) \geq 0.949$
3. $P(pfd_{sys} < 1.008 \times 10^{-5}) \geq 0.94$
4. $P(pfd_{sys} < 1.99 \times 10^{-7}) \geq 0.9$

Notice that even though the result 2 has a stronger claim than 3, it still has a higher confidence. This is because all these confidence bounds are conservative rather than exact values, and the degree of conservatism can vary from case to case.

4 Epistemic uncertainty: confidence bounds for *$pnqp_B$*

To use the theorems in the previous section an assessor needs to provide numerical values for his beliefs about the parameters *pfd_A* and *$pnqp_B$*, expressed in terms of (some of) percentiles (confidence bounds), means, variances.

It is well-known how to do this for *pfd_A*, for example based on evidence from operationally representative statistical testing: see, e.g., (Littlewood and Wright 1997). In this section we shall consider inference about *$pnqp_B$*. We shall restrict ourselves initially to the problem of finding a *confidence bound*, rather than mean or variance, for *$pnqp_B$* so that we can use Theorem 6 of the previous section to obtain a confidence bound for the system *pfd*.

We begin by recalling the urn model of Section 2. The development of the present program is here treated as the random selection of a program (ball) from the *B* urn. Programs in this urn have different *pfd*s (some will be zero), so the selection of a program is also a selection of a *pfd_B*. That is, the outcome of the selection is a random variable. Denote by *$f_B(p)$* the distribution of these *pfd_B*s. If this distribution were known, an assessor could compute *$pnqp_B$* (or even *pnp_B* to use in the LP theorems).

Note that this is an *objective* distribution: it is a property of urn *B*. Extending the thought experiment of Section 2 slightly, we could imagine, for each program randomly selected from urn *B*, executing *n* randomly selected operational demands. The proportion of these that fail, as *n* becomes infinite, is the *pfd* of that program. Now imagine doing this for many independently randomly selected *programs* from *B*, and forming a histogram of the different *pfd*s. As the number of programs increases, this converges to the distribution *$f_B(p)$*.

Unfortunately, of course, *$f_B(p)$* will be unknown: an assessor would be uncertain about this distribution, and thus about *$pnqp_B$*. It is this epistemic uncertainty that we wish to capture in a confidence bound for *$pnqp_B$*.

The reader may think that a way forward at this stage would be to assume that *$f_B(p)$* is a member of a parametric family of distributions – for example a Beta(α, β) family. In that case the epistemic uncertainty concerns solely the values of the parameters α, β. In principle an assessor could obtain a Bayesian (subjective) posterior distribution for (α, β) in the usual way (based, presumably, upon evidence concerning the nature and quality of the development process used to obtain *B* programs). From this he could compute a posterior distribution for $pnqp_B$ and thus his required confidence bounds (percentiles of that distribution). However, it does not seem reasonable to expect that an assessor could ever be *certain* that the distribution $f_B(p)$ was from the Beta (or any other) parametric family: this seems a very strong assumption that would be hard to justify.

We proceed, therefore, to present an example involving much more limited assumptions about the form of the unknown $f_B(p)$:

$$P(pfd_B = 0) = \theta \tag{26}$$

$$P(pfd_B > y) = \alpha \tag{27}$$

$$P(0 < pfd_B \leq \varepsilon) = \beta \tag{28}$$

i.e. essentially three percentiles only: see Figure 1.

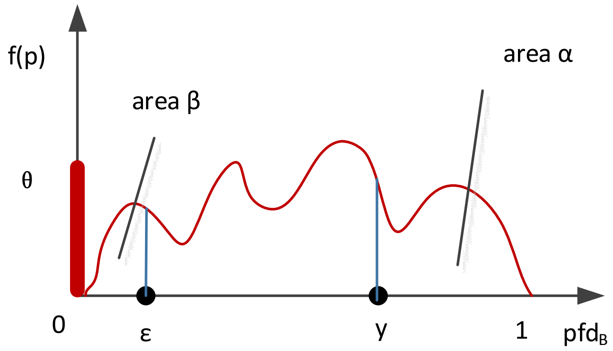

Figure 1 An example showing a hypothetical $f_B(p)$ distribution satisfying the constraints (26-28).

Of course, by making only very restricted assumptions like these we do not completely characterize a distribution for the *pfd*s from urn *B*: there will be an infinite number of distributions that satisfy (26) to (28). Our approach in what follows will be to choose the worst case distribution – i.e. the one that gives the most conservative results – in the spirit of our earlier work (Bishop, Bloomfield et al. 2011).

Here θ, β, α are unknown parameters, and it is about these that the assessor will eventually have to express his subjective beliefs: we have reduced the problem of epistemic uncertainty to this problem. We shall treat y and ε as given (i.e. known to the assessor). In particular ε is *not* a parameter about which an assessor will express prior belief, rather it can be regarded as a *requirement* for the definition of quasi-perfection here. If pfd_B is smaller than ε the assessor regards the system to be, informally, "effectively indistinguishable from perfect". The value required for ε will thus be calculated by the assessor: see footnote 2 for a hypothetical nuclear example.

We stress that this is only a single example of such a restricted set of assumptions for this stage of the reasoning. There are clearly other ways of doing this. There will be two main requirements in choosing such assumptions:

- it needs to be feasible for an assessor to obtain evidence to support quantitative beliefs about the unknown parameters;

- the conservatism that the assumptions induce in the final results, concerning system reliability, should not be so great as to render these results useless.

Readers may ask why we still need the parameter θ, representing probability of perfection, here in this quasi-perfection model. To recap: with the very limited prior beliefs in Figure 1, we have shown that evidence from failure-free operation (or operational testing) does not support an increase in probability of perfection that would be useful for assessing *system* reliability. However, we have shown here that this evidence, given the same prior beliefs, improves probability of *quasi*-perfection – and this can then be used in system assessment.

Note that just having a prior confidence in quasi-perfection (together with the other confidence bound in Figure 1) would also not support this kind of reasoning: the evidence would not help to improve confidence in quasi-perfection.

From Figure 1, we can see that the probability of quasi-perfection, pqp_B, is $\theta+\beta$, i.e. $pnqp_B=1-(\theta+\beta)$. Consider now the situation in which n demands have been executed without failure. Denoting by pqp_B^* the new (conditional) probability of quasi-perfection, and using Bayes Theorem, we have:

$$pqp_B^* = P(0 \le pfd_B \le \varepsilon \,|\, n \text{ failure-free demands}, f_B(p))$$

$$= \frac{\theta + \int_{0+}^{\varepsilon} (1-p)^n f_B(p)\,dp}{\theta + \int_{0+}^{\varepsilon} (1-p)^n f_B(p)\,dp + \int_{\varepsilon+}^{y} (1-p)^n f_B(p)\,dp + \int_{y+}^{1} (1-p)^n f_B(p)\,dp} \qquad (29)$$

We can interpret this in terms of our urn-based thought experiment of Section 2 as follows. We begin with urn B containing programs with *pfd*s having a distribution $f_B(p)$. We now take a vector of n randomly selected demands and execute every program in the urn on these demands. Some programs will survive all demands, some will fail on one or more demands. Remove all the latter from the urn. The remaining programs will have *pfd*s with some distribution $f_B^*(p)$, different from $f_B(p)$. In fact

$$f_B{}^*(p) = \frac{(1-p)^n f_B(p)}{\int_0^1 (1-p)^n f_B(p)\,dp} \qquad (30)$$

Then pqp_B^* is the new probability of quasi-perfection from this new distribution, concerning the "n-demand survivor" programs.

Notice that the Bayesian reasoning here – from pqp_B to pqp_B^* - concerns *this program and its ilk* (i.e. n-demand survivor programs from urn B). It does not involve any Bayesian inference about the *model*, represented by $f_B(p)$: i.e. we are not learning about the parameters θ, β, α that characterise this model.

Among all the distributions $f_B(p)$ that satisfy the assumed constraints above, (26) to (28), we can find the most conservative one, i.e. the one that makes the probability of quasi-perfection, pqp_B^*, as low as possible. It turns out (see proof in the Appendix), that this is the distribution with four points of support in Figure 2.[4]

[4] Strictly speaking, there may be other distributions that satisfy the constraints and give the same probability of quasi-perfection; but there are none that give a smaller value.

Using this conservative distribution, we obtain the most conservative – i.e. a lower bound on – probability of quasi-perfection:

$$pqp_B{}^* \geq \frac{\theta + (1-\varepsilon)^n\beta}{\theta + (1-\varepsilon)^n\beta + (1-\varepsilon)^n(1-\theta-\alpha-\beta) + (1-y)^n\alpha}$$

$$= \frac{\theta + (1-\varepsilon)^n\beta}{\theta + (1-\varepsilon)^n(1-\theta-\alpha) + (1-y)^n\alpha} \qquad (31)$$

$$\geq \frac{\theta}{\theta + (1-\varepsilon)^n(1-\theta-\alpha) + (1-y)^n\alpha} \qquad (32)$$

$$= 1 - G(\theta, \alpha) \text{ say,}$$

so that

$$pnqp_B^* < G(\theta, \alpha) \qquad (33)$$

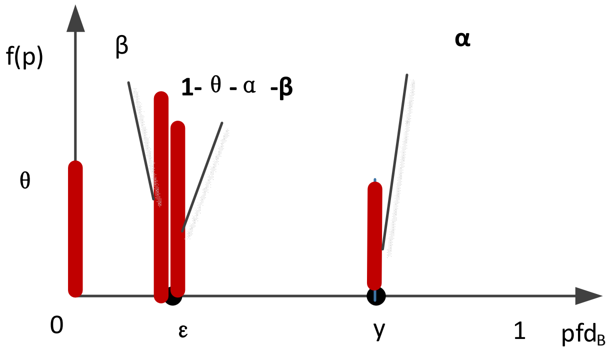

Figure 2 This four-point distribution gives the smallest posterior probability of quasi-perfection, subject to the prior constraints (26), (27), (28). There are four points of support here, but note that the mass *β* and the mass 1-*θ*-*α*-*β* are coincident at *ε* (strictly speaking the worst case is a limit for the *pfd* value associated to the point mass 1-*θ*-*α*-*β* tending to *ε* from above).

We give both these results, (31) and (32), because putting *β*=0 in (31) introduces the opportunity to simplify the assessor's task, albeit at the price of further conservatism: using (32) reduces the problem of inference to just two parameters, *θ*, *α*. Of all the values in (31) that *β* might take (i.e. the assessor might believe), *β*=0 is the most conservative: it gives the smallest probability of quasi-perfection. We shall use (32) rather than (31) in what follows.

The remaining problem is to obtain a (assessor's subjective, posterior) distribution for the unknown parameter vector (*θ*, *α*). From this bivariate distribution we could obtain confidence bounds for $pqp_B{}^*$ (or, equivalently, $pnqp_B{}^*$) using (32).

Because of the impracticality of eliciting bivariate prior beliefs, from which to compute a bivariate posterior distribution for (*θ*, *α*), we introduce another simplification – again at the expense of further conservatism – in the spirit of similar results in Section 2 and LP (Littlewood and Povyakalo 2013). In the following theorem we obtain conservative confidence bounds for $pnqp_B^*$ based only on *marginal* confidence bounds for the unknown parameters θ and α.

Theorem 7

If

$$P(\theta < z_\theta) = D_1 \tag{34}$$

and

$$P(\alpha < z_\alpha) = D_2 \tag{35}$$

then

$$P(pnqp^* \le G(z_\theta, z_\alpha)) \ge 1 - (D_1 + D_2) \tag{36}$$

(see Appendix for proof)

We thus have a confidence bound for the probability that channel B (having survived n demands without failure) is *nqp*, which is what is required, for example for theorem 6 of Section 3. Using Theorem 6 and Theorem 7 we are now able to obtain a conservative confidence bound for pfd_{sys}.

5 Confidence bounds for pfd_{sys}

Theorem 6 gives a conservative confidence bound for the system *pfd* in terms of a confidence bound on the *pfd* of channel A and a confidence bound on the *pnqp* of channel B. Theorem 7 gives the latter (conservatively) in terms of confidence bounds on θ and α. Combining these results we have:

Theorem 8

Given

$$P(pfd_A < P_A) = 1 - \alpha_A$$

and

$$P(\theta < z_\theta) = D_1$$

and

$$P(\alpha < z_\alpha) = D_2$$

we have

$$P\left(pfd_{sys} < \varepsilon \times \left(1 - G(z_\theta, z_\alpha)\right) + P_A \times G(z_\theta, z_\alpha)\right) > 1 - (\alpha_A + D_1 + D_2)$$

(37)

(see Appendix for proof)

Example 7

If we set $P_A=10^{-4}$, $\alpha_A=0.05$ and $D_1=D_2=0.05$ we obtain the results in Table 1.

We should emphasise that the numbers used here are merely illustrative, and intended only to give the reader a feel for how all this works. They are not meant to be realistic, in the sense that they would relate to any real system[5].

The table shows that confidence in quasi-perfection, and in a small system *pfd*, both grow as n increases – see columns seven and eight. However, this growth is modest for the kinds of values of n that might realistically be obtained in practice. Stronger claims about θ, i.e. a bound on z_θ of 0.9 in column 1, seem to make this growth in confidence greater.

We have not yet had the opportunity to investigate in detail how different values of the several parameters here influence the results. We leave this for future work. However, Table 2, shows an example of the advantage of this quasi-perfection approach over the "pure perfection" approach of our earlier work, reported in (Zhao 2015).

z_θ	ε	z_α	y	n	$G(z_\theta, z_\alpha)$, bound on $pnqp_B^*$	Bound on pfd_{sys} in left-hand side of equation (37)	Minimum confidence level: right-hand side of (37)
0.2222	0.000001	0.07407	0.001368	1	0.7777773251	7.8E-05	0.85
0.2222	0.000001	0.07407	0.001368	10^3	0.7646349293	7.67E-05	0.85
0.2222	0.000001	0.07407	0.001368	10^5	0.7413143312	7.439E-05	0.85
0.2222	0.000001	0.07407	0.001368	10^7	0.0001437642	1.014E-06	0.85
0.9	0.000001	0.07407	0.001368	1	0.0999087675	1.089E-05	0.85
0.9	0.000001	0.07407	0.001368	10^3	0.0473597461	5.689E-06	0.85
0.9	0.000001	0.07407	0.001368	10^5	0.0254031346	3.515E-06	0.85
0.9	0.000001	0.07407	0.001368	10^7	0.0000013078	1E-06	0.85

Table 1 The 6th column gives the bound on $pnqp_B^*$ in (36), and the confidence level here is at least 90%. The 7th and 8th columns give, respectively, the upper bound on pfd_{sys} and the minimum confidence in that bound.

If we compare the third and last rows of Table 2, based on our earlier "pure perfection" work, with the quasi-perfection results of the fourth and eighth rows of Table 1, we can see the improvement the latter brings. The bounds on pfd_{sys} in each of these rows

[5] In fact the numbers we used were loosely based on those obtained in the experiment by Knight and Leveson. They had a "population" of 27 versions which could be treated informally as our urn B. By extensively testing the members of this population, an estimate could be made of the distribution of pfd_B. In the 10^7 tests, 6 of the 27 versions had no failures. A point estimate of θ is 6/27=0.222222. Because we cannot be sure this is the true value, we treat it as a 95% lower confidence bound, i.e. z_θ=0.222222 and D_1=0.05 in equation (34). As we say above, we do not claim that this informal use of the Knight and Leveson results gives numbers that one would encounter in real-life applications.

of Table 2 are worse than the quasi-perfection equivalents of Table 1 – by almost two orders of magnitude in the first case, by a factor of more than two in the second case.

z_θ	ε	z_α	y	n	Bound on pnp_B^*	Bound on pfd_{sys} in LR result	Minimum confidence level
0.2222	0	0.07407	0.001368	10^3	0.764810913	7.64811E-05	0.85
0.2222	0	0.07407	0.001368	10^5	0.760025056	7.60025E-05	0.85
0.2222	0	0.07407	0.001368	10^7	0.760025056	7.60025E-05	0.85
0.9	0	0.07407	0.001368	10^3	0.047388938	4.73889E-06	0.85
0.9	0	0.07407	0.001368	10^5	0.028004277	2.80043E-06	0.85
0.9	0	0.07407	0.001368	10^7	0.028004277	2.80043E-06	0.85

Table 2 Results using the same parameter values as Table 1, but based on ε=0, i.e. pure perfection approach of (Zhao 2015).

We do not, of course, claim that the new model will *always* be superior to the older one in this way – in fact this is shown in other comparisons between the tables, where the perfection model can be superior to the quasi-perfection one. It seems that it is in cases of larger *n* (e.g. $n > 1/\varepsilon$) where there is benefit from using the quasi-perfection models. In fact, Table 2 suggests that this may be because there is a strong law of diminishing returns operating for the pure perfection models. Very large numbers of failure-free runs hardly improve the results: e.g. when increasing *n* from 10^5 to 10^7 the results remain the same (within the numerical accuracy of the table). In contrast, for the quasi-perfection models, similar increases in *n* produce significant improvements. We plan to investigate such issues in more detail in further work.

Readers should note, however, that from a practical point of view an assessor does not *need* to know *a priori* which of the different approaches will give superior results for his application. For his given value of ε (obtained, as we have suggested, from a wider safety case) and a particular value of *n* (i.e. the evidence that has been collected), he can apply *both* the pure *and* the quasi-perfection models and use the better of the two results. Whichever model this comes from, the result is guaranteed to be conservative.

6 Summary, further work and discussion

6.1 Summary of this modelling

The work reported here extends that reported earlier, particularly in (Littlewood and Rushby 2012, Littlewood and Povyakalo 2013, Zhao 2015). It addresses the problem of assessing the reliability – in fact probability of failure on demand (*pfd*) – for a demand-based one-out-of-two system, such as is commonly used for safety protection in many industries, including nuclear reactor protection.

The technical problems addressed here concern *aleatory and epistemic uncertainty*, particularly difficult issues of *dependence.*

1. *Aleatory dependence.* It is well-known that so-called “independently developed” channels of a fault-tolerant 1-o-o-2 system cannot be assumed to fail independently of one another. This means that the system *pfd* cannot be obtained simply by multiplying the channel *pfd*s. In the earlier LR work, in which one channel was “possibly perfect”, we showed that the product $pfd_A.pnp_B$ could be used as a conservative bound for system *pfd*. In the present work we generalize this result using a concept of *quasi-perfection* (*qp*), and prove a simple new conservative bound (Section 2) involving pfd_A, $pnqp_B$ and the computed parameter ε that defines *qp*. This generalizes the LR result.
2. *Epistemic dependence.* The parameters of this model, pfd_A, $pnqp_B$, will be unknown. A formal Bayesian treatment requires an assessor to describe his uncertainty about them in terms of a bi-variate distribution. It is well-known that assessors find it difficult, if not impossible, to describe their uncertainty as a complete distribution – and particular difficulty is associated with *dependence* of their beliefs about the two parameters. In Section 3 we prove several theorems that allow claims for system *pfd* to be made in terms of only *marginal* claims for the parameters – i.e. knowledge of epistemic dependence is not required. All these results are guaranteed to be conservative.
3. *Limited assessor knowledge.* Assessors are unlikely to be able to state even their marginal beliefs in terms of complete distributions. The theorems of Section 3, then, involve only different kinds of *limited* beliefs: percentiles, means, variances. The price paid here for simplifying the assessor’s task – eliminating dependence and using only limited beliefs – is further conservatism. These theorems of Section 3 generalise the results of LP.
4. *Confidence bounds for* $pnqp_B$ The results of Sections 2 and 3 reduce the problem to one of Bayesian statistical inference – based on evidence such as extensive testing – about the parameters pfd_A, $pnqp_B$ treated separately. For the first of these, the solutions are well known and simple. For $pnqp_B$ things are not so simple, and Section 4 provides a solution, again requiring from the assessor only limited knowledge. Once again, the results are guaranteed to be conservative.
5. *Final percentile claim for* pfd_{sys}. In Section 5 we sketch how all this can be used to make a conservative top-level claim for pfd_{sys}, expressed as a percentile.

In summary, the modeling reported here presents an end-to-end assurance argument about the *pfd* of a 1-o-o-2 system. In doing so it responds to the real difficulties encountered in such arguments, namely issues of dependence and limited assessor prior belief.

An important point is that, over the several stages of the end-to-end assurance argument, *conservatism is guaranteed.* One way of looking at this is that conservatism is “the price paid” to avoid these traditional difficulties of dependence and limited knowledge. Putting it more positively, a guarantee of conservatism is an advantage – indeed one could say a necessity – for safety claims about the kinds of safety-critical systems for which such arguments will be used. We are not aware of other ways of achieving results that are guaranteed to be conservative in this way. We shall expand on these comments briefly in the discussion in Section 6.3.

Before continuing to discuss possible further work, we would like to make clear a subtle distinction between the work reported here (and that in the earlier LR/LP papers), and previous work on a similar problem reported in (Zhao 2015). In (Zhao

2015), we were concerned with uncertainty about an *event*, “system not perfect”. The model we presented there allows an assessor to “learn” about the probability of this event: i.e. we showed how the assessor’s subjective prior probability for this event became, after collecting evidence, a posterior probability for the event. In contrast, in LR/LP and here, we are dealing with (epistemic) uncertainty about “in the world,” objective *parameters* – respectively, *probability not perfect* and *probability not quasi-perfect*. We deal here with subjective beliefs about the numerical values of these parameters – and obtain posterior beliefs about them.

6.2 Proposed further work

At various points in the preceding account we have suggested that further work is needed. Below we list some of the most important.

- *Where do assessors get beliefs from?* This is a question that is always asked about Bayesian analysis – particularly *prior* beliefs. The classical response is that assessors must *have* prior beliefs that they bring to the problem, and the only issue is how to “elicit” these accurately. Unfortunately, this is often not the case, and we believe applications to safety-critical systems, as here, may be such a situation. Furthermore, priors, like any other component of an argument, need to be formulated by those making a claim but argued as justifiable to those vetting the argument. We expect it will be possible to give some guidance, for example, about plausible priors for θ from the literature and industrial experience on effectiveness of verification methods in use, and experience of failure-free operation of systems that have seen massive exposure (cf discussion in (Strigini and Povyakalo 2013)).
- *Other kinds of evidence.* Section 4 only uses evidence of failure-free working. In practice other kinds of evidence are available and should be used. Notable examples include: evidence from verification (see, e.g. (Littlewood and Wright 2007)) and evidence of process quality based on past product experience. These will typically feed into the priors available *before* the operational testing. We have begun looking at formal models of uncertainties in verification, and of the contribution that could be made by knowledge of the efficacy of the development processes used (e.g. using operational experience from previously developed “similar” systems).
- *“What if?” calculations.* We have illustrated our results throughout with numerical examples, and one of these is “complete” in the sense that it results in a confidence bound for the system *pfd*. The numbers used in the examples are, of course, chosen just to illustrate our approach – in particular to show what is needed to populate our whole end-to-end argument numerically. We have not so far attempted any substantial “what if?” calculations – e.g. to see the relative effects of different aspects of assessor knowledge upon the final claim.
- *Worked examples based on realistic numbers.* We need to see what kinds of results we get from our modeling when using numbers that are as close to realistic as we can make them. One question we would like to investigate concerns *how* conservative our results are in such situations: some may be too

conservative to be useful[6]. We expect that the specific evidence available about a specific system will determine which one, among various ways of supporting system *pfd* claims, will give the most favourable, but *still conservative*, result. We intend to interact with the sponsors of this research to see whether we can use numbers that are as close to realistic – in their view – as possible.

- *Other kinds of system claim; "lifetime measures".* Our worked example here concerns only a confidence bound for the system *pfd*, i.e, using only Theorem 6 of Section 3. We would also like to work examples using the other theorems of Section 3, involving expected values for system *pfd*. All the results in the present paper concern the probability of system failure on (*one*) demand in one way or another. In practice, a wider safety case might be concerned with ensuring a high probability of no failures throughout the life of a system, or fleet of systems – i.e. over an expected number of demands, rather than a single one. Our results do not currently address such questions, but we see opportunities for extension in this direction.
- *Choice of simplifying assumptions in Section 4.* As we stated earlier, the assumptions we use here, (26) to (28), are clearly not the only ones that we could have used. To complete the reasoning for this particular end-to-end argument, it will be necessary for an assessor to obtain the bounds in (34) and (35): the feasibility of doing this (i.e. obtaining the empirical evidence to support such claims) will be a test of the practicality of the approach outlined here. It seems worthwhile seeking other approaches, however – i.e. other simplifying assumptions for this stage of the argument. There are two aims in such a quest: that the parameters involved can be estimated from available real-world data; that the necessary simplification involved does not induce excessive conservatism on the final results.
- *Choice of* ε. There is more than one feasible approach to select the value of ε, the *pfd* bound that defines "quasi perfection". One may select a value ε such that the target probability of experiencing no failures over the lifetime of the system is satisfied. Alternatively, considering that there are always trade-offs between confidence *bounds* and confidence *levels*, one can instead, with some additional numerical or algebraic calculations, choose ε such as to get the most favourable claim feasible, within the constraint of required conservatism.

6.3 Discussion

We have emphasized at various points in this paper that we have produced a complete "end-to-end" argument in support of claims for the *pfd* of a 1-o-o-2 system, taking account of all the different kinds of uncertainty. The different stages of the argument each involve conservative assumptions that provide, importantly, a *guarantee* that the final results – claims about the system *pfd* – are conservative. This chain of conservatism may, of course, mean that the results from the overall argument are *very* conservative. This seems to be so in the case of the illustrative example we show in Section 5, but it may not be so for a different selection of parameter values, or for

[6] Of course, even if this is the case, one could respond by saying "so be it", and challenge an advocate of a particular system, who regards our results as *too* pessimistic, to proffer alternative ways of supporting their claims.

different simplifying assumptions from the ones in Section 4. For this reason we propose to conduct the "what if?" exercise described above.

Our "complete conservatism" approach in the presence of all kinds of uncertainty is, we believe, novel. In particular, we are not aware of any evaluations of real diverse systems that are argued in such a way that a claim of complete conservatism can be made convincingly. Such evaluations typically deal with *pfd* claims for *both* channels, in contrast to our approach involving perfection or quasi-perfection for one of the channels. They thus have to deal with the problem of failure dependence between channels, which precludes the simple claim that $pfd_{sys}=pfd_A.pfd_B$. Ways around this difficulty tend to be rather informal.

In fact there are cases where channel failures have simply been assumed to be independent, and $pfd_A.pfd_B$ used for system *pfd.* Typically, justification of independence in such cases is not believable; indeed, we do not know credible means of claiming independence. Making such an assumption can, of course, be dangerously optimistic.

We have seen an argument that recognizes the problem of failure dependence between channels and deals with it along the following lines:

> "Our claims for the *pfd*s of the two channels are themselves very conservative: we know that each system is much better than the numbers pfd_A and pfd_B that we are claiming. So, when we use the product $pfd_A.pfd_B$ for the system *pfd* in our safety case, we can be sure that *this* is conservative."

The problem here, of course, is that this is a comparison between apples and oranges: how can we be certain that the "levels of conservatism" in each of the *pfd* claims are enough to counter the "level of dependence" between the channel *A* and channel *B* failures? We know of no way to reason about a trade-off between such different things so that this claim can be supported rigorously.

Of course, our approaches in this report and in LR, LP (Littlewood and Rushby 2012, Littlewood and Povyakalo 2013), seem likely to bring their own problems. In the first place, the reader will have seen that the treatment of epistemic uncertainty here (particularly in Section 4), and in LP, is not easy. And of course the full end-to-end argument may give *very* conservative results.

A perceptive reviewer of an earlier version of this paper asked whether even the simplified demands upon the assessor that our new approach entails would be simple enough in practice:

> [The reviewer questioned whether] "…the expectation that even the simplified reliability argument demonstrated here is any more realistic for use in practice than any of the previous ones. The paper's central premise is that limitations in assessor's prior knowledge about failure probability make it impractical to express it as complete distributions. Rather, assessors are asked to speculate about marginal values of such distributions. This appears simpler, indeed. But when faced with legal consequences of making such assumptions, is the authors' "expectation" that assessors would be any more willing to speculate on marginal values of the same distributions they were unwilling/unable to describe? The consequences of errors in judgment would be the same."

As we say above, we accept that the assessor is still faced with a difficult task. However, we maintain that the task *is* considerably simplified by our approach.

So Theorem 1 provides a conservative bound that is not available in the classical case where claims are made about *pfd*s for both channels: the difficult problem of aleatory dependence (between channel failures) has been eliminated.

Then, instead of the need to specify a complete bivariate distribution, the assessor needs to specify (for example) only a couple of *percentiles* of *univariate* marginal distributions. Eliminating the need for any assessment of epistemic dependence, in particular, is an important simplification.

The assessor has a complete guarantee that in reducing the problem in this way the results will be conservative. In fact, in providing only (say) percentile information a nervous assessor has an opportunity to introduce further conservatism – something that would be much harder to do if they were required to specify a complete distribution.

We believe that the beliefs required of an assessor here (e.g. a couple of percentiles) are the very minimum needed to provide useable claims about the system reliability. In fact one view of some of our results here is that they operate within a region lying between prior beliefs that are "too minimal to be useful" and ones that are "too demanding of an assessor to realistically expect them to be trustworthy."

Having said all that, we acknowledge that sometimes an assessor may *not* be able to provide even these minimal beliefs. If that is the case, then so be it. It suggests, we believe, that they are not in a position to make trustworthy claims about the reliability of the system in question using the reasoning proposed here. In such a case, "refusing to speculate" (in our reviewer's terminology) would be an assessor's safe and honest option.

We think our approach – in particular its guaranteed end-to-end conservatism – provides a rigorous formalism for assessing these kinds of systems. It could be used to challenge the trustworthiness of more informal approaches when these result in less conservative claims.

In fact readers might ask whether we need our full end-to-end treatment in order to get useful results that are at least an improvement on the classical approach involving pfd_A and pfd_B.

Indeed, we often see reliability models used assuming that the parameter values are correct; a use that may be justified when there is extremely high confidence that they are very close approximations to the truth or that they are worst-case bounds. This simplification, applied to our approach, would work out as follows. Given point estimates of pfd_A and pnp_B, we would treat each of these as "true" and use a simple product of them as a bound on system *pfd*. Whilst ignoring epistemic uncertainty about the parameters in this way is "wrong", it is nevertheless superior to using the naïve product of pfd_A and pfd_B. Specifically, it ignores *only* epistemic uncertainty; the other approach ignores epistemic uncertainty *and* failure dependence. That is, if we had correct values for pfd_A and pnp_B[7] our result *would* be a bound on expected system *pfd*. The same is not true of the second approach: even with true values for pfd_A and pfd_B their product is not such a guaranteed bound on system *pfd*.

[7] We emphasise again, as earlier in this section, that we *do* here have a notion of "true value" for these parameters. They are "in the world" parameters – i.e. they are not subjective beliefs about the probabilities of events in the world. They have unknown values, of course, and our subjective beliefs – and Bayesian analysis – centre upon these values.

A slightly less crude argument is similar – but superior – to the one in quotes above. Suppose we are prepared to believe that the channel claims, pfd_A and pnp_B, are each *conservative with certainty*, as is reasoned above. In that case their product is, with certainty, a conservative bound on expected system *pfd*. That is, because of the basic LR result, we do not need to trade off apples and oranges as in the quoted example.

The reader may think this kind of reasoning, ignoring epistemic uncertainty completely, is *too* crude. Whilst not defending it, we just remark that such a "plug in numbers" approach is quite common in the more classical approaches to these problems.

But we can go further along our conservative chain of reasoning: i.e. we can use less crude approximate arguments that still avoid some of the problems of our full end-to-end approach. Suppose, for example, we were prepared simply to use the theorems of Section 3 without taking advantage of the methods of Section 4 for refining confidence about pnp_B. The resulting system *pfd* claims will have taken some account of epistemic uncertainty, albeit not as much as the full end-to-end argument. We claim, somewhat tentatively, that reasoning like this is superior to anything we have seen in arguments based upon claims of *pfd* for both channels, where epistemic uncertainty is ignored completely or treated very informally. At least here we can claim that if the numbers we plug in to Section 3 theorems are true, then the resulting system *pfd* claim is guaranteed to be conservative.

The point here is that our end-to-end argument involves several stages, and there is guaranteed conservatism at each stage. This means that the conclusion of the full end-to-end argument is guaranteed to be conservative, as we have said, but it also means that if we stop before completing all stages, we still have conservatism in what we can claim.

Curtailing the full end-to-end treatment of uncertainty in ways like this would make the task of the assessor very much easier, but at the price of not ending up with a *guaranteed* conservative claim for the system *pfd* – because some uncertainty will not be accounted for[8]. Nevertheless, it might be argued that even though not perfect, such reasoning is superior to that used in examples such as the ones quoted above.

Appendix

Proof for theorem 2

Denote the unknown joint probability $P(pfd_A > \alpha_A, pnqp_B > \alpha_B)$, i.e. lying in BCEF in Figure A1, by z.

$$\mathrm{E(Pfd}_{sys}) \le E[\varepsilon \times (1 - pqnp_B) + pfd_A \times pqnp_B]$$

$$= E[\varepsilon + (pfd_A - \varepsilon) \times pnqp_B]$$

$$\le [\varepsilon + (P_A - \varepsilon) \times P_B] \times (1 - \alpha_A - \alpha_B + z) + [\varepsilon + (P_A - \varepsilon)] \times (\alpha_B - z)$$

[8] Or, putting it another way, it places on the assessor a requirement to be *certain* of the values of certain parameters. With such certainty, there *will* be guaranteed conservatism in the conclusions.

$+[\varepsilon+(1-\varepsilon)\times P_B]\times(\alpha_A-z)+\mathrm{z}$
(A1)

$$=\ \varepsilon+(P_A-\varepsilon)P_B+(1-P_A)P_B\alpha_A+(P_A-\varepsilon)(1-P_B)\alpha_B+(1-P_A)(1-P_B)z \qquad \text{(A2)}$$

which yields the result since the z coefficient here is positive, and by definition z cannot exceed α_A or α_B.

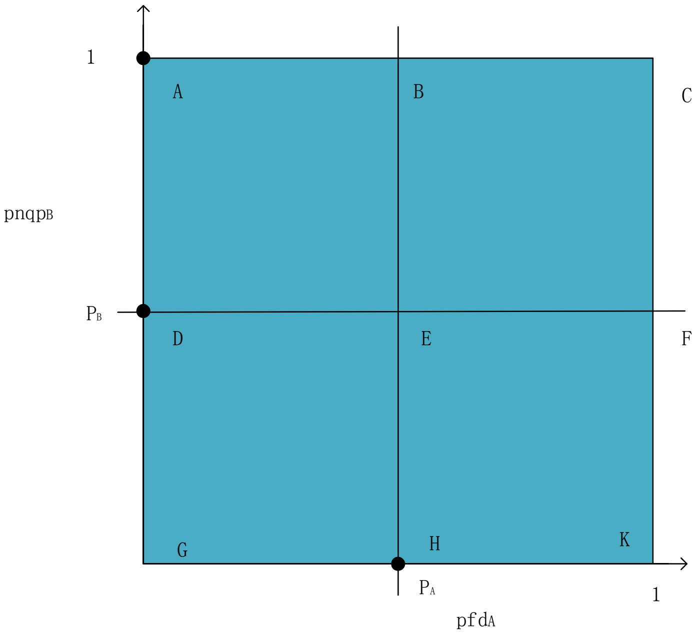

Figure A1

Expression (A1) can be explained as follows. Within the area DEGH, as $\mathrm{pfd_A} \leq \mathrm{P_A}$, and $pnqp_B \leq \mathrm{P_B}$, the product $\varepsilon+(pfd_A-\varepsilon)\times pnqp_B$ is a random variable which is everywhere smaller than the value at the top righthand corner $\varepsilon+(P_A-\varepsilon)\times P_B$. And the probability associated with the area DEGH is $(1-\alpha_A-\alpha_B+z)$. Thus the contribution to system mean *pfd* associated with DEGH is bounded above by the product $[\varepsilon+(P_A-\varepsilon)\times P_B]\times(1-\alpha_A-\alpha_B+z)$, which is the first term in (A1). Similarly, within the area ABED, the product $\varepsilon+(pfd_A-\varepsilon)\times pnqp_B$ is a random variable which is everywhere smaller than $\varepsilon+(P_A-\varepsilon)\times 1$, and the probability associated with the area ABED is (α_B-z). So the contribution to the mean system *pfd* of this rectangle is bounded by $[\varepsilon+(P_A-\varepsilon)]\times(\alpha_B-z)$, which is the second term. The same reasoning applied on the BCEF and EFHK gives the whole result (A1), the form (A2) of which is then obtained by collecting coefficients of the three doubt-related parameters $\alpha_A,\ \alpha_B, z$.

QED

Proof for theorem 3

The Proof here is similar to the previous one, but in terms of the four rectangles in QSWG.

$\mathrm{E(Pfd}_{sys}) \leq E[\varepsilon\times(1-pqnp_B)+pfd_A\times pqnp_B]$

$= E[\varepsilon + (pfd_A - \varepsilon) \times pnqp_B]$

$$= [\varepsilon + (P_A - \varepsilon) \times P_B] \times (1 - \alpha_A - \alpha_B + z) + [\varepsilon + (P_A - \varepsilon) \times P_B^U] \times (\alpha_B - z) + [\varepsilon + (\mathrm{P}_A^U - \varepsilon) \times P_B] \times (\alpha_A - z) + [\varepsilon + (P_A^U - \varepsilon) \times P_B^U] \times \mathrm{z}$$

$$= \varepsilon + (P_A - \varepsilon)P_B + (P_A^U - P_A)P_B\alpha_A + (P_A - \varepsilon)(P_B^U - P_B)\alpha_B + (P_A^U - P_A)(P_B^U - P_B)\mathrm{z}$$

from which the bound follows just as for Theorem 2.

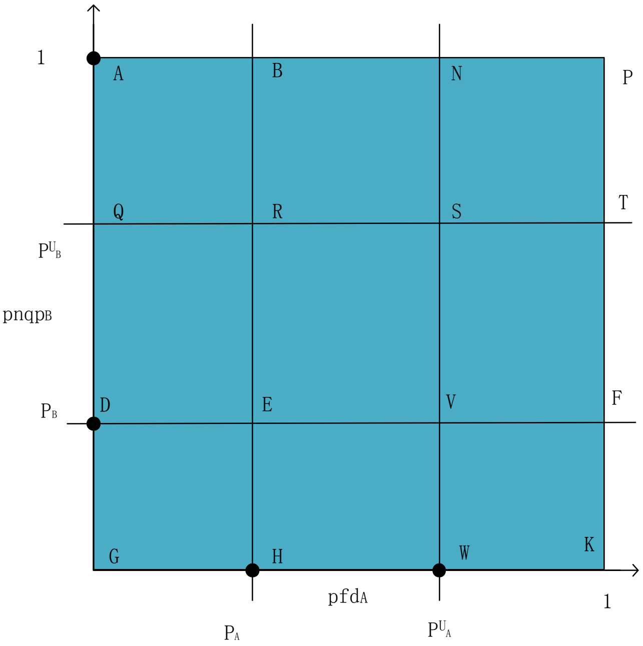

Figure A2

The reasoning about the DEHG is exactly the same with previous one. For the rectangle QRDE, is bounded by $[\varepsilon + (P_A - \varepsilon) \times P_B^U] \times (\alpha_B - \mathrm{z})$; the contribution from EVWH is bounded by $[\varepsilon + (\mathrm{P}_A^U - \varepsilon) \times P_B] \times (\alpha_A - z)$; that from RSVE is bounded by $[\varepsilon + (P_A^U - \varepsilon) \times P_B^U] \times \mathrm{z}$.

Notice that, when $P_A^U = 1$ and $P_B^U = 1$ the result here reduces to the result of theorem 2.

QED

Proof for theorem 4

By the Cauchy–Schwarz inequality,

$$\left(E(pfd_A \times pqnp_B)\right)^2 \leq E\left(pfd_A{}^2\right)E(pqnp_B{}^2)$$
$$= \left(E(pfd_A)^2 + Var(pfd_A)\right)\left(E(pqnp_B)^2 + Var(pqnp_B)\right)$$

And

$E(pfd_A)^2 + Var(pfd_A) < (E(pfd_A) + SD(pfd_A))^2$, with a similar expression involving pnqpB.

So,

$E(pfd_{sys}) \leq E[\varepsilon \times (1 - pqnp_B) + pfd_A \times pqnp_B]$

$$= \varepsilon - \varepsilon \times E(pqnp_B) + E\left(pfd_A \times pqnp_B\right)$$

$$< \varepsilon - \varepsilon \times E(pqnp_B) + \sqrt{\left(E(pfd_A)^2 + Var(pfd_A)\right)\left(E(pqnp_B)^2 + Var(pqnp_B)\right)}$$

$$< \varepsilon - \varepsilon \times E(pqnp_B) + (E(pfd_A) \ + SD(pfd_A))(E(pfd_A) \ + SD(pfd_A))$$

QED

Lemma

If $0 \leq X \leq 1,\ p > 0,\ P(X > p) = \alpha$ and $E(X) \leq p\ \leq \alpha$, then

$$E(X^2) \leq\ (1 + p)E(X) - \alpha p \leq p^2 + p(1 - \alpha)$$

Proof for Lemma

The proof is based on the representation of the distribution of *X* as a mixture of two scaled versions of random variables *U* and *V,* distributed, respectively, within the intervals *U*∈[0, 1] and *V*∈(0,1]. Suppose that *X* has cdf F_X, and let *U, V* and *Z* be three mutually independent random variables, the first two continuous with respective marginal distributions:

$F_U(u) = \frac{F_X(pu)}{1-\alpha}$; $F_V(v) = \frac{F_X(p+(1-p)v)\ -\ (1-\alpha)}{\alpha}$; and the third *Z* a Bernoulli distribution with $P(Z = 1) =\ \alpha$.

Define, in terms of these three, a random variable

Y = (1-Z) p U + Z [p+(1-p)V].

It is easily verified that *Y is then distributed identically to X.* Furthermore, $Y^2 = (1 - Z)p^2\, U^2\ +\ Z\, [p^2 + 2p(1 - p)\, V + (1 - p)^2\, V^2\,]$. (For example, one may square the expression for *Y* and use the relations *Z(1-Z)=0, Z=Z*^{2}, *1-Z =(1-Z)*^{2} satisfied by the Bernoulli variable *Z*.)

We will use this construction to obtain expressions for the first two moments of the common distribution of *X* and *Y,* relying as we proceed on the mutual independence of *(U,V,Z)* to factorize the expectations of any product terms. Firstly

$$E(X) =\ E(1 - Z)\, p\, E(U) + E(Z)[p + (1 - p)E(V)]$$

$$=\ \ (1 - \alpha)\, p\, E(U) + \alpha\, [p + (1 - p)E(V)]\ \ . \tag{A3}$$

Secondly,

$$E(X^2) = E[\,(1 - Z)p^2\, U^2\ \ +\ \ Z\, [\,p^2 + 2p(1 - p)\, V + (1 - p)^2\, V^2\,]\ \ ]$$

$$=\ (1 - \alpha)\, p^2\, E(U^2) + \alpha\, [p^2 + 2p(1 - p)\, E(V) + (1 - p)^2\, E(V^2)\,]$$

$$\leq\ (1 - \alpha)\, p^2\, E(U)\ + \alpha\, [p^2 + (1 - p^2)\, E(V)\,] \tag{A4}$$

because E(*U*^{2}) $\leq E(U)$, E(*V*^{2}) $\leq E(V)$. Subtracting (1+p) times (A3) from (A4) eliminates *E(V)*, and rearranging the result of this gives

$$E(X^2) \leq (1 + p)E(X) - (1 - \alpha)p\, E(U) - \alpha p$$

$$\leq (1 + p)E(X) - \alpha p$$

$$\leq p^2 + p(1 - \alpha) \tag{A5}$$

making use, at the last line, of our original assumption, $E(X) \leq p$.

QED

Remarks on Attainability of the Upper Bound in this Lemma

If $p = \alpha$, then the upper bound (A5) of this Lemma is attained when $E[V(1 - V)] = E(U) = 0$, i.e. when P(X = 0) = 1 – α and P(X = 1) = α.

If $p < \alpha$, then the upper bound (A5) of this Lemma is "asymptotically" attained when $E[V(1 - V)] = E(U) = 0$. For example suppose, for some $0 < \delta' < \frac{p(1-\alpha)}{\alpha(1-p)}$,

$P(U = 0) = 1$,

$$P(V = \delta') = \frac{1 - E(V)}{1 - \delta'} \ ;$$

$$P(V = 1) = \frac{E(V) - \delta'}{1 - \delta'} \ ;$$

$$E(V) = \frac{p(1 - \alpha)}{\alpha(1 - p)} \ ,$$

So that the associated *X* then has the discrete three-point distribution,

$$P(X = 0) = 1 - \alpha;$$

$$P(X = p + \ \delta' - \delta' p) = \alpha \frac{1 - E(V)}{1 - \delta'} = \frac{\alpha - p}{(1 - \delta')(1 - p)}$$

$$P(X = 1) = \alpha \frac{E(V) - \delta'}{1 - \delta'} = \frac{p(1 - \alpha) - \alpha(1 - p)\delta'}{(1 - \delta')(1 - p)} \quad ; \delta' > 0; \delta' \rightarrow 0.$$

To check that the bound is approached, write δ for $\delta'(1 - p)$ so as to simplify a little to

$$P(X = 0) = 1 - \alpha;$$

$$P(X = p + \ \delta) = \frac{\alpha - p}{1 - p - \delta}$$

$$P(X = 1) = \frac{p(1 - \alpha) - \alpha\delta}{1 - p - \delta} \quad ; \delta > 0; \delta \rightarrow 0.$$

One may confirm directly that this three-point, δ-paramterized distribution has *E(X)=p* and

$$E(X^2) = p^2 + p(1 - \alpha) - \delta(\alpha - p)$$

which approaches our upper bound from below as $\delta \rightarrow 0+$.

Proof for theorem 5

In accordance with the lemma:

$$E\left(pfd_A{}^2\right) \leq (1 + P_A)E(pfd_A) - \alpha_A P_A \leq P_A^2 + P_A(1 - \alpha_A)$$

and

$$E(pnqp_B^2) \leq \ (1 + P_B)E(pqnp_B) - \alpha_B P_B \leq P_B^2 + P_B(1 - \alpha_B).$$

Then, by the Cauchy–Schwarz inequality,

$$E(pfd_A \times pqnp_B) \leq \sqrt{E\left(pfd_A{}^2\right)E(pqnp_B{}^2)} \leq$$

$$\sqrt{(1+P_A)E(pfd_A)-\alpha_A P_A} \times \sqrt{(1+P_B)E(pqnp_B)-\alpha_B P_B} \leq$$
$$\sqrt{P_A^2+P_A(1-\alpha_A)} \times \sqrt{P_B^2+P_B(1-\alpha_B)}$$

So,

$$E(pfd_{sys}) \leq E[\varepsilon \times (1-pqnp_B)+pfd_A \times pqnp_B]$$
$$= \varepsilon - \varepsilon \times E(pqnp_B) + E(pfd_A \times pqnp_B)$$
$$\leq \varepsilon - \varepsilon \times E(pqnp_B) +$$
$$\sqrt{(1+P_A)E(pfd_A)-\alpha_A P_A} \times \sqrt{(1+P_B)E(pqnp_B)-\alpha_B P_B}$$
$$\leq \varepsilon - \varepsilon \times E(pqnp_B) + \sqrt{P_A^2+P_A(1-\alpha_A)} \times \sqrt{P_B^2+P_B(1-\alpha_B)}$$

QED

Proof for theorem 6

First, the contour plot of the expression $\varepsilon(1-y)+x\,y$ is as figure A3

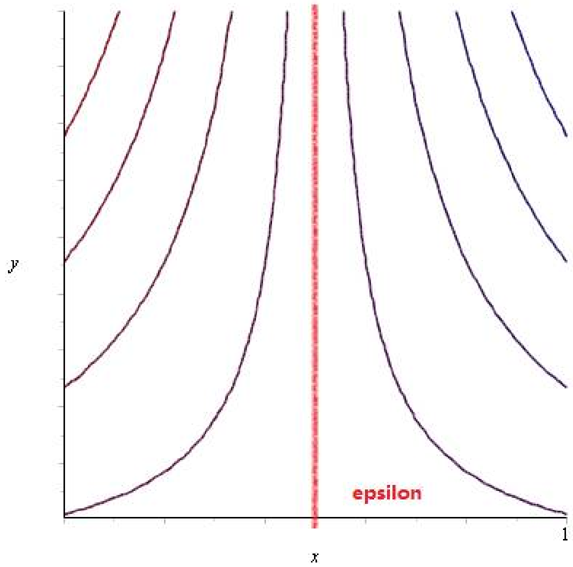

Figure A3

As

$$pfd_{sys} \leq \varepsilon \times (1-pnqp_B)+pfd_A \times pnqp_B$$

So

$$P\big(pfd_{sys} < \varepsilon \times (1-P_B)+P_A \times P_B\big) \geq P(\varepsilon \times (1-pnqp_B)+pfd_A \times pnqp_B < \varepsilon \times (1-P_B)+P_A \times P_B)$$

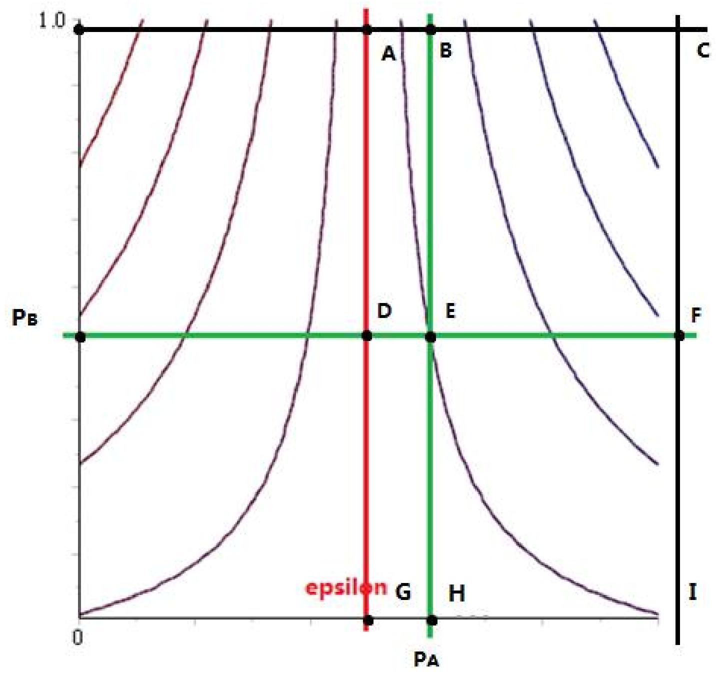

Figure A4

From the figure A4, we could have

$$P(\varepsilon \times (1 - pnqp_B) + pfd_A \times pnqp_B > \varepsilon \times (1 - P_B) + P_A \times P_B)$$

$$< P(pfd_A > P_A) + P(pnqp_B > P_B)$$

This is because the left hand side is the probability mass associated with the area above the hyperbola (the one goes through E point) in Figure A4. This is contained in the union of the two rectangular parts of the unit square respectively above the horizontal green line and to the right of the vertical green line, whose probabilities are $(pfd_A > P_A)$ and $P(pnqp_B > P_B)$.

Then,

$$P\big(pfd_{sys} < \varepsilon \times (1 - P_B) + P_A \times P_B\big)$$

$$\geq P(\varepsilon \times (1 - pnqp_B) + pfd_A \times pnqp_B < \varepsilon \times (1 - P_B) + P_A \times P_B)$$

$$= 1 - P\big(\varepsilon \times (1 - pnqp_B) + pfd_A \times pnqp_B > \varepsilon \times (1 - P_B) + P_A \times P_B\big)$$

$$> 1 - (\alpha_A + \alpha_B)$$

QED

Proof of most conservative $f_B^*(p)$ in section 4

$$pqp_B^* = P(0 \leq pfd \leq \varepsilon \mid n\, failure\, free\, tests, f_B(p))$$

$$= \frac{P(pfd = 0)P(n\, failure\, free\, tests|pfd = 0) + P(0 < pfd \leq \varepsilon)P(n\, failure\, free\, tests|0 < pfd \leq \varepsilon)}{P(n\, failure\, free\, tests)}$$

$$= \frac{\theta + \int_{0+}^{\varepsilon}(1-p)^n f_B(p)dp}{\theta + \int_{0+}^{\varepsilon}(1-p)^n f_B(p)dp + \int_{\varepsilon+}^{y}(1-p)^n f_B(p)dp + \int_{y+}^{1}(1-p)^n f_B(p)dp}$$

By the mean value theorem for integrals, we could find 3 values, say P_1 P_2 and P_3, satisfying the equations below,

$$(1-P_1)^n\int_{0+}^{\varepsilon} f_B(p)dp = \int_{0+}^{\varepsilon}(1-p)^n f_B(p)dp$$

$$(1-P_2)^n\int_{\varepsilon+}^{y} f_B(p)dp = \int_{\varepsilon+}^{y}(1-p)^n f_B(p)dp$$

$$(1-P_3)^n\int_{y+}^{1} f_B(p)dp = \int_{y+}^{1}(1-p)^n f_B(p)dp$$

Where,

$$0 < P_1 \le \varepsilon, \varepsilon < P_2 \le y \; and \; y < P_3 \le 1$$

And from the prior constraints (26), (27) and (28) in section 4,

$$\int_{0+}^{\varepsilon} f_B(p)dp = \beta, \int_{\varepsilon+}^{y} f_B(p)dp = 1-\theta-\alpha-\beta \; and \int_{y+}^{1} f_B(p)dp = \alpha$$

Then our objective function turns to:

$$\begin{aligned} pqp_B^* &= \frac{\theta+(1-P_1)^n\beta}{\theta+(1-P_1)^n\beta+(1-P_2)^n(1-\theta-\alpha-\beta)+(1-P_3)^n\alpha} \\ &= \frac{1}{1+\dfrac{(1-P_2)^n(1-\theta-\alpha-\beta)+(1-P_3)^n\alpha}{\theta+(1-P_1)^n\beta}} \end{aligned}$$

To minimize it is to maximize P_1 and minimize P_2 and P_3. That is let $P_1=\varepsilon$, $P_2=\varepsilon^+$ and $P_3=y^+$, which is in response to the prior distribution in Figure 2. Then,

$$\begin{aligned} pqp_B^* &\ge \frac{\theta+(1-\varepsilon)^n\beta}{\theta+(1-\varepsilon)^n\beta+(1-\varepsilon)^n(1-\theta-\alpha-\beta)+(1-y)^n\alpha} \\ &= \frac{\theta+(1-\varepsilon)^n\beta}{\theta+(1-\varepsilon)^n(1-\theta-\alpha)+(1-y)^n\alpha} \\ &\ge \frac{\theta}{\theta+(1-\varepsilon)^n(1-\theta-\alpha)+(1-y)^n\alpha} \end{aligned}$$

QED

Proof for Theorem 7

From the main text of the paper, we have:

$$pnqp_B^* \le G(\theta,\alpha)$$

Now $G(\theta,\alpha)$ is a decreasing function of both θ and α, so:

$$P\big(\mathrm{pnqp}_B^* \le G(z_\theta,z_\alpha)\big) \ge P(\theta \ge z_\theta, \alpha \ge z_\alpha)$$

The joint distribution of θ and α satisfies:

$$\begin{aligned} P(\theta \ge z_\theta,\alpha \ge z_\alpha) &= 1-P(\theta < z_\theta)-P(\alpha < z_\alpha)+P(\theta < z_\theta,\alpha < z_\alpha) \\ &\ge 1-P(\theta < z_\theta)-P(\alpha < z_\alpha) = 1-D_1-D_2 \end{aligned}$$

So

$$P(\mathrm{pnqp}_B^* \leq G(z_\theta, z_\alpha)) \geq 1 - D_1 - D_2$$

QED

Proof for Theorem 8

From the proof of theorem 6, we know that

$$P\left(pfd_{sys} < \varepsilon \times \left(1 - G(Z_\theta, Z_\alpha)\right) + P_A \times G(Z_\theta, Z_\alpha)\right) \\ \geq 1 - (\alpha_A + P(pnqp_B^* \geq G(\theta, \alpha)))$$

by replacing P_B with $G(z_\theta, z_\alpha)$ and α_B with $P\left(\mathrm{pnqp}_B^* \geq G(\theta, \alpha)\right)$

Using the theorem 7 result, we have:

$$P\left(pfd_{sys} < \varepsilon \times \left(1 - G(Z_\theta, Z_\alpha)\right) + P_A \times G(Z_\theta, Z_\alpha)\right) \geq 1 - (\alpha_A + D_1 + D_2)$$

QED

References

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 (2013). Statistical Summary of Commercial Airplane Accidents, Worldwide Operations, 1959-2012. Seattle, Aviation Safety, Boeing Commercial Airplanes.

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.

FAA (1988). Advisory Circular 25.1309-1A: System design and analysis. Washington DC, Federal Aviation Administration.

HSE (2011). Step 4 Control and Instrumentation Assessment of the EDF and AREVA UK EPR Reactor. Bootle, Health and Safety Executive, Office for Nuclear Regulation.

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.

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 bounds for the pfd of a 1-out-of-2 software-based system based on an assessor's subjective probability of "not worse than independence"." IEEE Trans Software Engineering **39**(12): 1641-1653.

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.

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

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. (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.

Acknowledgements

We thank the three reviewers of an earlier version of this paper for their careful readings and insightful comments; these have helped us improve the presentation of this material considerably.

Xingyu Zhao's work reported here was supported by a PhD Studentship funded by City University London.

The work of the other authors was partly supported 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.