Bounds on survival probability given mean probability of failure per demand; and the paradoxical advantages of uncertainty

February 2014

Abstract

When deciding whether to accept into service a new safety-critical system, or choosing between alternative systems, uncertainty about the parameters that affect future failure probability may be a major problem. This uncertainty can be extreme if there is the possibility of unknown design errors (e.g. in software), or wide variation between nominally equivalent components.

We study the effect of parameter uncertainty on future reliability (survival probability), for systems required to have low risk of even only one failure or accident over the long term (e.g. their whole operational lifetime) and characterised by a single reliability parameter (e.g. probability of failure per demand — *pfd*). A complete mathematical treatment requires stating a probability distribution for any parameter with uncertain value. This is hard, so calculations are often performed using point estimates, like the expected value.

We investigate conditions under which such simplified descriptions yield reliability values that are sure to be pessimistic (or optimistic) bounds for a prediction based on the true distribution. Two important observations are: (i) using the expected value of the reliability parameter as its true value guarantees a *pessimistic* estimate of reliability, a useful property in most safety-related decisions; (ii) with a given *expected pfd*, broader distributions (in a formally defined meaning of "broader"), that is, systems that are *a priori* "less predictable", lower the risk of failures or accidents.

Result (i) justifies the simplification of using a mean in reliability modelling; we discuss within which scope this justification applies, and explore related scenarios, e.g. how things improve if we can test the system before operation. Result (ii) offers more flexible ways of bounding reliability predictions, but also has important, often counter-intuitive implications for decision making in various areas, like selection of components, project management, and product acceptance or licensing. For instance, in regulatory decision making dilemmas may arise in which the goal of minimising risk runs counter to other commonly held priorities, like predictability of risk; in safety assessment using expert opinion, the commonly recognised risk of experts being "overconfident" may be less dangerous than their being *under*confident.

Contents

Some Notation and Abbreviations Used

A, B labels denoting the two components of a simple series or parallel composite system

Ap. appendix

cdf cumulative distribution function of a random variable

dem. demands

E expectation operator as applied to random variable

f_Q probability density function of *pfd*

i, I lower case i is density function of leftward-moved mass in broadening operation of section 5.1; upper case I is an interval containing all of this mass before it is moved left, under-barred to denote its infimum (left-hand endpoint), and over-barred to denote its supremum (right hand endpoint)

j, J lower case j is density function of leftward-moved mass in broadening operation of section 5.1; upper case J is an interval containing all of this mass before it is moved right, under-barred to denote its infimum (left-hand endpoint), and over-barred to denote its supremum (right hand endpoint)

k ratio of both sizes of the two masses moved apart and also of their respective distances moved, so that mean is preserved in broadening operation of section 5.1

$\Lambda, \lambda, \lambda^*$ Continuous-time failure rate parameter: as random variable, instantiated value, or mean – analogously to Q below

MTTF mean time to failure

pfd probability of failure per demand of component or system

Pr probability

PRA probabilistic risk assessment

Q, q, q^* *pfd*, upper case Q when regarded as an uninstantiated random variable; or lower case q to denote a particular realised value; or starred q^* to denote its mean value $E(Q)$

$R(t)$ reliability function R evaluated at usually discrete time t demands. $R(t)$=Pr(no failure occurs over first t demands); sometimes with further semicolon-separated parameter arguments. Also used for analogous continuous time reliability

surv. survives

$\mathcal{S}$ a system subject to discrete demands on each of which it may succeed or failure

t system operating time, usually discrete (number of demands $t = 0, 1, 2, \ldots$) unless otherwise stated

Thm. theorem

U $1 - Q$, where Q is the *pfd* as a random variable

w.r.t. with respect to

1 Introduction

Predictions of reliability and safety through probabilistic modelling depend on the values of model parameters, e.g. component failure rates, which are often uncertain.

The main application scenario that motivates our research involves decisions on accepting a software product for use in a safety critical application requiring low accident probability over the operational life of the system in which it is embedded. For instance, an explicit requirement in civil aviation is that “catastrophic failure conditions” be “so unlikely that they are not anticipated to occur during the entire operational life of all airplanes of one type” [16]. Nuclear power protection systems may have required or claimed *pfd* bounds like 10^{-7} or 10^{-9} [23, 22], to assure low probability of even one failure during operational life. Formally, the system's predicted reliability function, concerning those failures that may cause accidents, must be close to 1 at the end of the intended operational life.

With software, decisions are made especially difficult by uncertainties about whether design faults are present, and about their effect on probability of failure. The same difficulty arises with respect to the probability of any system failures due to design faults. Similar decision problems may also arise regarding physical failures of components, if there is a concern about broad variation in reliability parameters, as for instance with the current alarm about electronic component supplies "contaminated" with unreliable "counterfeit" components.

Our reference scenario is a system $\mathcal{S}$ with high required confidence of operating until the end of its service life without failure causing accident. Subject to discrete demands, $\mathcal{S}$'s failure process is completely characterised by a constant *probability of failure per demand* (*pfd*). Examples are failure of software due to design faults (the original motivation of our work), or hardware without aging or maintenance.

Mathematically similar scenarios exist regarding reliability even without safety implications, e.g. when a component should last for the lifetime of the system of which it is part, because it cannot be replaced or repaired, by either design (as in many consumer products) or necessity (e.g. in spacecraft).

We must predict $\mathcal{S}$'s probability of surviving t future discrete, independent demands – its reliability $R(t)$ in discrete time – with t an upper bound on the lifetime number of demands, if accident-free.[1] This would be straightforward except for uncertainty about the *pfd* value [3], arising e.g. because *pfd* is:

- inferred from reliability databases on components that are similar, but not identical, to the one for which a prediction is sought, and/or that operate in potentially different conditions, affecting their reliability differently. If the details of which systems failed and when are missing or not released;
- guessed using indirect evidence, as e.g. often done for *pfd*s due to software design faults.

This uncertainty can in theory be rigorously described by a subjective probability distribution for the value of each parameter. However, an assessor has seldom a clear idea of this distribution, and many calculations are *de facto* performed by treating their available $E(pfd)$ estimate as though it were the true *pfd*. Sensitivity analysis may be used to check that small variations in the estimate only cause acceptable prediction error, but in practice much reasoning among practitioners only deals with a point estimate $E(pfd)$, without acknowledging that, in fact, the shape of the probability distribution of the *pfd* may also have a substantial effect on the predicted value sought (e.g. reliability over a given period of operation), and this effect may be non-obvious.

Thus, using a point *pfd* estimate to calculate a system's lifetime survival probability may lead to errors of various kinds [1, 9].

Uncertainty about parameter values is a typical case of *epistemic* uncertainty in predictions (i.e., uncertainty arising from lack of knowledge rather than from an "inherent randomness" of the process studied). Epistemic uncertainty is widely studied [19, 20] and many formal mathematical methods have been proposed for dealing with it, but much normal practice does not use them. The practical approaches, e.g. in the nuclear industry [13, 15], are essentially of two kinds: qualitative criteria for accepting evidence (e.g., requiring that parameter value be derived from evidence that is more clearly pertinent to the specific plant, the more critical the parameters in question are) and numerical methods for performing either sensitivity analysis or calculations taking into account the complete probability distributions that describe uncertainty on the parameters. To cite [13]:

> Because the impact of parameter uncertainty can be addressed in terms of a probability distribution on the numerical results of the PRA, it is straightforward to compare a point value, be it the mean, the 95th percentile, or some other representative value with an

[1] This number of demands is usually a random variable, but we will treat the problem with reference to a fixed t. Conclusions for a random number of demands can be derived if required.

> acceptance guideline or criterion . . . For most regulatory applications, that value is specified to be the mean [. . .] The mean values referred to are the arithmetic means of the probability distributions that result from the propagation of the uncertainties on the input parameters.

Uncertainty propagation methods will in theory produce accurate results for any given distribution; but their application is hard: apart from computational complexity, their fundamental drawbacks are in delivering numerical results rather than insight on how the various aspects of parameter uncertainty may affect the results, and in requiring a complete description of the parameters' distribution, which in practice may be hard to specify with any degree of soundly based consensus (especially if we consider that the uncertainties on the various parameters are not statistically independent – a concern called sometimes "epistemic correlation"). When the problem is to extract a distribution for the parameters in question from detailed failure data about multiple similar systems, typical approaches to modelling and inference use *hierarchical* or *empirical Bayes* [3, Ch.8],[42, Ap.A],[11],[40, §6],[31].

These approaches reason with particular, analytically convenient, typically conjugate, *parametric families* of *pfd* distributions for the reliability parameter (e.g. failure rate) of a specific system; the former positing an (updateable) prior distribution over the chosen parametric family, and the latter applying one or a combination of established statistical inference techniques to select a preferred family member.

But in many application areas the issue of parameter uncertainty is not yet generally addressed [9]. Thus, in this paper we seek simpler, albeit less general, methods for reasoning with epistemic uncertainty in a simple and common scenario. We characterise the errors incurred when making assumptions on distributions of which only the *mean* is given. We observe results simple enough for routine application, and yet unexpected enough (from anecdotal evidence: our own discussions with researchers and practitioners) that we must examine their implications for common decision making scenarios.

In the rest of this paper, section 2 discusses causes of parameter uncertainty and consequent prediction errors. Next follow some mathematical results: section 3 states upper and lower bounds on the reliability after t demands, if only the *expected* system *pfd* is known; section 4 presents some extensions and consequences, including the case of reliability in continuous time; section 5 shows that the bounds of section 3 are a special case of more general reliability bounds, whereby, given the expected system *pfd*, a "broader" distribution of this parameter (in a specific meaning of "broader") implies higher reliability; section 6 discusses to what extent these theorems can be extended to multiple-parameter reliability functions. We then proceed in section 7 to discuss various implications for practical decision making, including simplifications of certain decision problems, but also difficulties, including a need to specify in detail the criteria one wishes to apply for accepting systems into operation, or for ranking alternative systems; we offer examples of counter-intuitive implications of the theorems and dilemmas they may create. Section 8 is a summary of conclusions.

2 Uncertainty on parameters and its effects

2.1 The inevitable uncertainty on the *pfd*

To estimate system $\mathcal{S}$'s *pfd*, one often relies on experience with similar systems: statistics that have been collected, or informal "expert judgement". These approaches can be seen as selecting a "reference" population of systems that are similar to system $\mathcal{S}$, and then using their estimated *pfd*s (usually their mean) to estimate the *pfd* for system $\mathcal{S}$. It is known that these systems are not identical, and they only define a statistical distribution of *pfd* values: system $\mathcal{S}$'s *pfd* may be anywhere in the range observed; or even outside it, although perhaps with low probability. The degree to which this process gives confidence varies. For software, arguments that the probability of failure is low thanks to a high quality development process – where "high quality" is a judgement based on experience – in practice use as "reference set" previous software products produced by ostensibly similar processes [30]. In the case with $\mathcal{S}$ known to be a "true" member of the reference population (e.g. a physical item, mass-produced, by the same factory, to the same design, in the same production batch) and having counted the total number of failures for many nominally identical systems, and the total number of demands over which the failures occurred, we have a point estimate for the mean *pfd* over the set of all the systems observed.

However, circumstances may make it infeasible to decide whether the inferred average describes a homogeneous population, in which the *pfd* is practically identical for all individuals, or how heterogeneous the population really is: whether some components will be substantially more, or less, reliable than this average, when used in the context of current interest. E.g., the historic failure rates observed may be low, the history of system instances (under representative operation) limited, environmental variation effects on historic data not well understood, or in other respects, our risk assessor's access to original data restricted, either because it is not saved, or is confidential.

Additional uncertainties arise because, for any system in our reference set, the *pfd* is estimated from observed demands and failures, so that rather than a measure known with certainty we have confidence intervals with an associated level of confidence. In many cases, furthermore, system $\mathcal{S}$ is known to differ from the reference set in some respect, but it is not clear how this dissimilarity should affect one's estimate of $\mathcal{S}$'s *pfd*. Alternatively, with copies of a system operating in different installations, there may be statistics of usage and failures that are sufficient to infer an average *pfd*, but also suggest that the true *pfd* varies among different installations, for unknown reasons. A subjective distribution of *pfd* of the system in a new installation needs then to include high uncertainty about the true value, with known mean.

All the above problems may be present when using data from reliability databases [17], and even more so in situations like software reliability assessment, which is often based on conformance to required development and verification practices, hoping that the faithful application of standard precautions offers a high enough confidence in a certain upper bound on the system's *pfd* [33, 34]. In practice, even the explicit statement of confidence is usually omitted when applying common standards that cover software safety (e.g. the popular IEC 61508 [24]). And yet cases of systems that are found *a posteriori* to be seriously less reliable than expected, despite rigorous assessment, are not uncommon. Complex systems that failed on first operational use include the early Space Shuttle, the Altona rail station control systems, the Ariane V [36]. In this last case, the probability of failure per demand (mission) was 1. That is, for complex systems subject to design faults, the subjective distribution of the *pfd* must allow it to be possibly very high, albeit with low probability. Therefore, trying to reason, conservatively, assuming a "worst case" *pfd* would not work: it would predict unacceptably low reliability even for a system that is likely to be quite reliable.

In the case of a software system, $\mathcal{S}$'s failures are due to systematic causes only, and one can test $\mathcal{S}$ for as long as affordable, with demands sampled from the distribution anticipated to occur in operation, without fear of wear and tear, instead of relying on its similarity with other systems (as when using reliability databases). Observing high reliability over an observation period or number of demands directly increases confidence of a low *pfd* for the specific system $\mathcal{S}$, and formal statistical inference tells us by how much [4, 33, 37, 35], as some combination of confidence bounds and confidence levels (in the case of classical inference) or a posterior probability distribution (with Bayesian inference).

2.2 Bayesian approach: using a *pfd* distribution

To reason rigorously about the effects of these uncertainties, the standard approach is Bayesian. The value of the *pfd* of system $\mathcal{S}$ is seen as a random variable Q, with probability density function $f_Q(q)$. If $\mathcal{S}$ is sampled from a supply of nominally equivalent systems which actually have a range of different *pfd* values (due for instance to manufacturing tolerances), this subjective distribution for the specific $\mathcal{S}$ represents the statistical distribution of the parameter in the concrete population.

Given $f_Q(x)$, one can then derive the effects of this distribution on any prediction of interest, e.g. what is a 99% "Bayesian confidence bound" on the true *pfd*, i.e. that value q_{99} such that we have 99% probability of $\mathcal{S}$ being at least as good as q_{99}, that is, that $\int_0^{q_{99}} f_Q(x)\,dx = 0.99$, and, more importantly, predictions about future operation, for instance the probability of surviving t

future demands without failures:

$$Pr(\text{survival for } \geq t \text{ demands}) = \int_0^1 (1-x)^t f_Q(x)\,dx \tag{1}$$

It is clear that these measures depend on the distribution $f_Q(q)$. Yet, reliability calculations are often performed using $E(Q)$ as a single point "location" representative of this distribution f_Q, both to avoid complex computations, and also because f_Q may be difficult to confidently specify. Recognising that this may lead to errors in prediction, an attempt is often made to be conservative by using a *pessimistic estimate* of this expected value.[2] But it is still necessary to ask whether the degree of pessimism introduced in the estimate of the expected *pfd* is enough to compensate for the possible errors introduced by performing calculations on the basis of an expected *pfd*.

2.3 Errors in reliability prediction - probability of surviving *t* demands

Comparing two systems' probabilities of survival on the basis of their expected (mean) *pfd*s may be misleading. E.g., Fig. 1 refers to two hypothetical systems, $\mathcal{S}_1$ and $\mathcal{S}_2$, whose *pfd*s have Beta probability distributions with respective means 0.1 and 0.05. By focusing on the mean *pfd*, a decision maker could easily conclude that $\mathcal{S}_2$ is the system to choose. But Fig. 2 shows that actually $\mathcal{S}_1$, despite having the higher (worse) average *pfd*, is better than $\mathcal{S}_2$ for scenarios requiring survival of more than 32 demands.

The next graph, in Fig. 3, also shows the predictions that would be obtained by substituting the mean *pfd* for the full distribution. We see that if we approximate the two *pfd* distributions with their means, $\mathcal{S}_1$'s reliability indeed appears to be permanently worse than $\mathcal{S}_2$'s; but $\mathcal{S}_1$'s *real* reliability is much greater than that calculated using $\mathcal{S}_1$'s mean *pfd*, and is better than $\mathcal{S}_2$'s for $t > 32$. Using the mean *pfd* for comparison is indeed completely misleading.

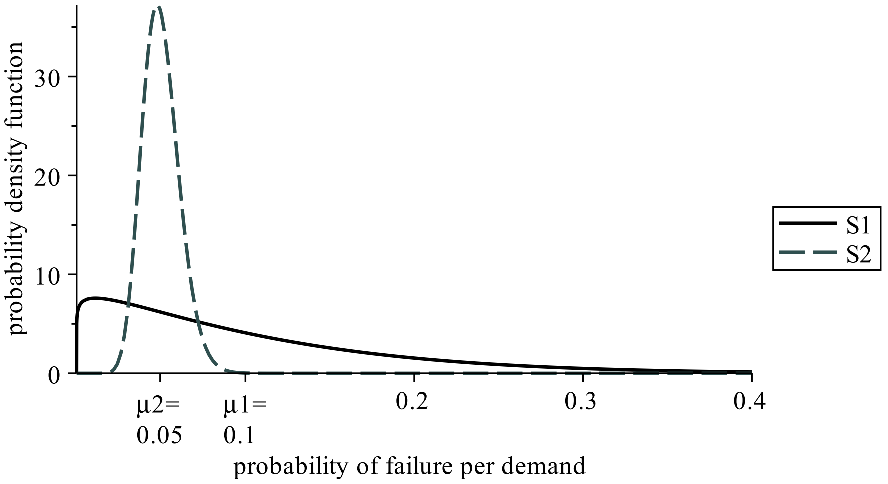


Fig. 1: Distributions (probability density functions) of the *pfd* for two hypothetical systems $\mathcal{S}_1$ and $\mathcal{S}_2$. Both are Beta distributions (probability density function $f(x) \propto x^{(a-1)}(1-x)^{(b-1)}$). $\mathcal{S}_1$ has mean *pfd* $\mu_1 = 0.1$, given by parameters of the Beta distribution $a_1 = 1.1$, $b_1 = 9.9$; $\mathcal{S}_2$ has lower mean *pfd*, $\mu_2 = 0.05$, but with parameters $a_2 = 20$; $b_2 = 380$, giving the narrower distribution shown.

Since a given mean *pfd* allows for a set of different reliability functions (probabilities of survival without failure)[3], using the mean *pfd* for predicting the reliability of one system may be as

[2] In other cases, there may be an explicit claim that the *pfd* is lower, *with certainty*, than an upper bound that guarantees the desired long-term reliability level, i.e. $q \leq q_{bound}$ with certainty and $(1-q_{bound})^t$, with t the lifetime number of demands, is acceptably large. This is not always practical, however.

[3] These reliability functions have different shapes even though they are all described by the simple geometric

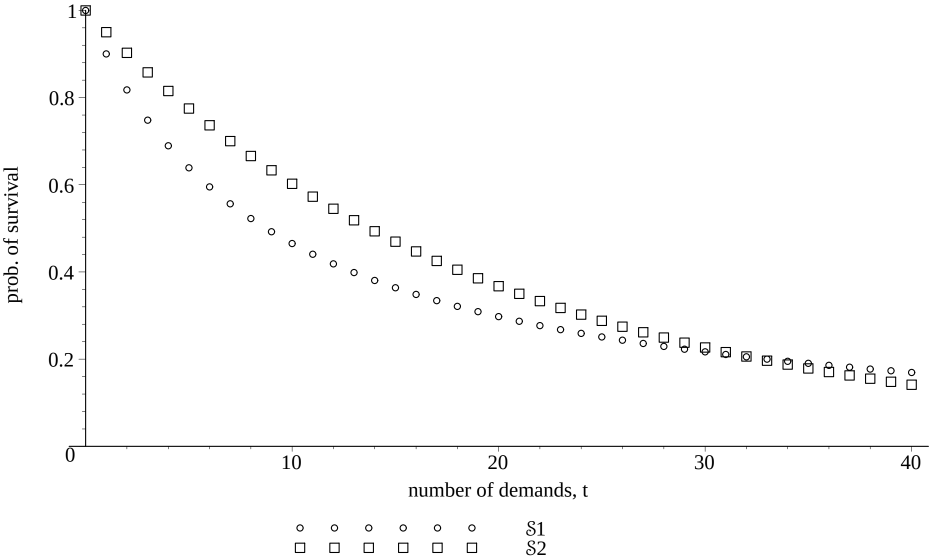


Fig. 2: Reliability functions (probability of surviving at least t demands – equation (1)) for two systems with the *pfd* distributions f_Q shown in Fig 1.

misleading as using it for comparisons between systems. Rather than giving further examples, we will demonstrate this by finding bounds on the possible reliability functions and thus on errors.

3 Bounds on reliability, given the mean *pfd*

Given that the *pfd* of a system is a random variable Q with an unknown probability density function $f_Q(x)$, but with known mean $E(q) = q^*$, the following properties hold regarding its probability of survival over t demands (their proofs are given in Ap. A):

THEOREM 1. *The* lowest *probability of survival (i.e., reliability), over all distributions of Q such that $E(Q) = q^*$, is equal to that obtained by assuming the system* pfd *to be equal to the mean of its distribution, q^**

$$Pr(survival\ for\ \geq t\ \ demands) = \int_0^1 (1-x)^t f_Q(x)\,dx \geq (1-q^*)^t$$

THEOREM 2. *The* highest *probability of survival (i.e., reliability), over all distributions of Q such that $E(Q) = q^*$, is equal to that obtained by assuming that the distribution of the system* pfd, *Q, is*

$$\Pr(Q{=}1) = q^*\,, \qquad \Pr(Q{=}0) = 1{-}q^*$$

$$\Pr(survival\ for\ \geq t\ \ demands) = \int_0^1 (1-x)^t f_Q(x)\,dx \leq 1-q^*$$

Note that these theorems state not only that the probability of survival for t demands, under the condition $E(Q) = q^*$, is bounded by $(1-q^*)^t$ and $(1-q^*)$, but that these two values are actually possible. This is so, because both the extreme distributions assumed in calculating the two bounds are actually legitimate probability distributions that satisfy the condition $E(Q) = q^*$. The assumptions for the two theorems correspond to the two discrete probability assignments shown in Fig. 4. So, we have the bounds:

function $(1-q)^t$, *conditional* on the value q of the *pfd*. Since the *pfd* may take different values with different probabilities, the resulting reliability function is a *mixture* of geometric reliability functions; not itself geometric.

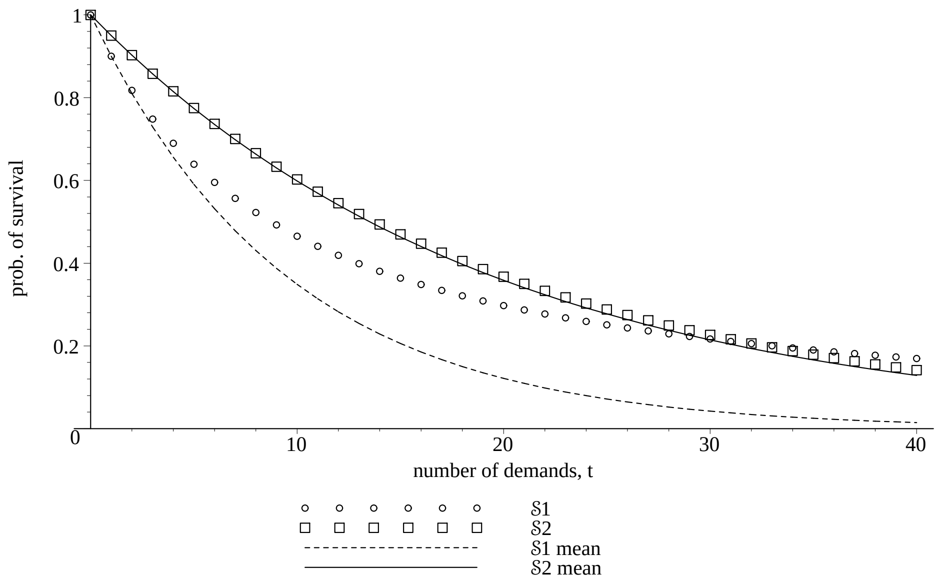


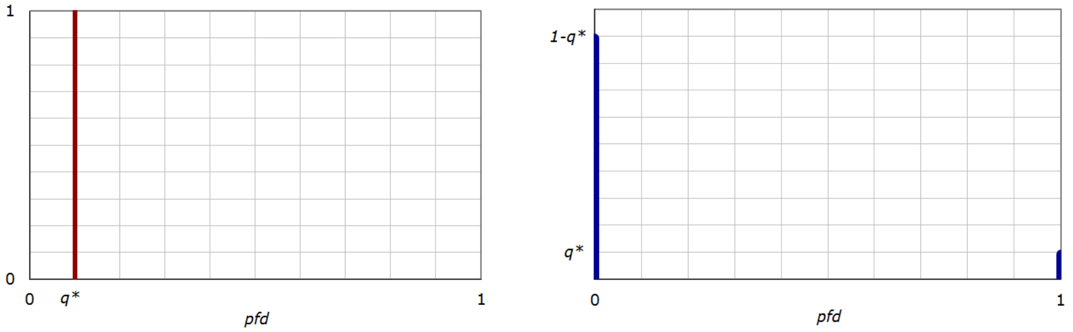
Fig. 3: The same two reliability functions as in previous figure, plus (as continuous lines for readability) reliability functions for hypothetical systems with *pfd* equal to the mean *pfd*s of $\mathcal{S}_1$ and $\mathcal{S}_2$. The curves labelled $\mathcal{S}_2$ and $\mathcal{S}_{2mean}$ are closer together than those labelled $\mathcal{S}_1$ and $\mathcal{S}_{1mean}$, which is intuitively explained by the probability distrib. of $\mathcal{S}_2$'s *pfd* being narrower than that of $\mathcal{S}_1$'s *pfd*. Also, the $\mathcal{S}_{2mean}$ curve lies completely below the $\mathcal{S}_2$ curve, and the $\mathcal{S}_{1mean}$ curve below the $\mathcal{S}_1$ curve, which holds in general by Thm. 1.

$$(1-q^*)^t \le \Pr(\text{survival for } \ge t \text{ demands}) \le 1-q^* \,, \tag{2}$$

while the common linear approximation $1-tq^*$ is a less tight lower bound on reliability.

Thm. 1 is based on Jensen's inequality [25, 10] and the convexity w.r.t. q of the geometric reliability function $(1-q)^t$ (strict convexity for $t>1$). Ap. A has a full proof.

An intuitive description of what these theorems state is as follows: any distribution of the *pfd* (for instance in terms of frequencies of occurrence in a population) implies a combination of individual items having greater *pfd* than the average, and thus higher probabilities of failing during the intended t-demand operational life, and others having lower *pfd* and accordingly lower probability of failing over the same life duration. The "best" of all these distributions has the highest frequency of items with 0 likelihood of ever failing. Given the known expected *pfd*, this distribution must also include a few items that would fail at every demand. But in a critical application, considering potentially catastrophic failures, what matters is whether a system fails



Worst case distribution (Thm. 1)

Best case distribution (Thm. 2)

Fig. 4: The two extreme *pfd* distributions. The graphs show discrete probability masses associated with values of *pfd*. They can be interpreted as representing probability density functions, with the thick vertical segments indicating Dirac's delta functions.

even only once in t demands. Whether a system would tend to fail once only or multiple times is immaterial, since the *first* failure will terminate the life of the item: even if the failure does not permanently disable it, it will cause its withdrawal from service. In the case of a design fault, e.g. in software, a dangerous failure will require a design change, applied to all copies, determining a new distribution of *pfd* and requiring a new acceptance decision.[4]

We are thus only interested in the system's reliability function having high values over the interval $[0, t]$, not in achieving a low MTTF.

3.1 Practical Implications of the Bounds on Reliability: conservatism

Thm. 1 states that using the mean *pfd* as if it were the true value guarantees pessimistic reliability predictions. In safety assessment, when the *pfd* is the probability of dangerous failure, this will often be good news: if approximation is inevitable, erring on the side of pessimism is more acceptable than erring on the side of optimism. Therefore, **using the mean *pfd* is a "safe" approximation**: the common recommendation to base acceptance decisions on comparing the mean with an acceptability threshold can be argued to be justified as conservative.

The bad news from Thm. 2 is that this error may be quite large. Indeed, for large values of t, the worst case reliability will be close to 0 while the best will be close to $1-q^*$. So, if the real distribution is quite broad, the error may exaggerate risk by orders of magnitude.

This approximation may cause us to discard systems that are indeed quite safe. Furthermore, it may cause us to rank the safety of different systems wrongly. As a striking example, a system with mean *pfd* 10^{-4}, with a very narrow distribution, will survive 1000 demands with probability $(1-10^{-4})^{1000} = 0.905$; while another system with mean $pfd = 10^{-2}$, but with a broad distribution approximating the "best case" of Thm. 2, will survive the same 1000 demands with much higher probability, close to 0.99: this latter system, which *"is 100 time worse on average"*, is almost *10 times less likely to fail* over the 1000 demands.

Next we study the extent of these possible errors.

3.2 Errors from using mean *pfd* for ranking alternatives

Are there any conditions in which only knowing the ordering between the mean *pfd*s of two systems also tells us the ordering between their reliability? Suppose we know that $q_1^* > q_2^*$. Is then $\mathcal{S}_2$ more reliable than $\mathcal{S}_1$: $R_1(t) < R_2(t)$? We observe that:

- this is certainly true for the *first* demand, $t = 1$;
- but even as early as the very next demand ($t = 2$), this ordering may be inverted and we could find $R_1(t) > R_2(t)$). To exclude this possibility, we would need to know that $(1 - q_2^*)^2$, the worst possible $R_2(2)$, is greater than $1 - q_1^*$, the best possible $R_1(2)$. Solving this inequality yields $q_1^* \geq q_2^*(2 - q_2^*) \approx 2q_2^*$: for "smaller mean *pfd*" to imply "better reliability at least until the second demand", the larger mean *pfd* needs to be at least *twice* as great as the smaller;
- what precedes applies for low t. In the long run, for larger t, the mean *pfd* becomes less relevant than the probability of *pfd* values "low enough" for likely survival for t demands. As $t \to \infty$, the reliability functions of $\mathcal{S}_1$ and $\mathcal{S}_2$ asymptotically tend (from above) to $Pr(pfd_1{=}0)$ and $Pr(pfd_2{=}0)$, whatever the other characteristics of the two *pfd* distributions (about the plausibility, or not, of non-zero $Pr(pfd = 0)$, *cf* subsection 7.4.3). So, given q_1^* (the greater of the two mean *pfd*s), a necessary condition for a "reversal of roles" (that is, $\lim_{t\to\infty} R_2(t) < \lim_{t\to\infty} R_1(t)$: $\mathcal{S}_1$ becoming eventually more reliable than $\mathcal{S}_2$) is $Pr(pfd_2{=}0) < 1 - q_1^*$, because we necessarily have $1 - q_1^* = E(1{-}pfd_1) \geq Pr(1{-}pfd_1 = 1)$ (simply because $1{-}pfd_1$ cannot be negative). Note that the two reliability functions will then cross at least once, as in Figure 2, since the reliability for $t = 1$ is exactly $1 - E(q)$, and thus higher for $\mathcal{S}_2$.

[4] Even if operation is resumed before these changes, it will be with extra precautions (*cf* e.g. the use of "Airworthiness directives" regarding flight-critical systems) that effectively change the circumstances of use, and thus the *pfd* distribution, and/or the reliability requirements. In any case, the previous reliability predictions become outdated after this first failure.

- to generalize for large but finite t, the smaller *pfd* only guarantees the higher reliability as long as, roughly, $t < q_1^*/q_2^*$, the ratio between the greater and the smaller mean *pfd*.
  We show this by comparing the lifetime failure probabilities, via the ratio $(1 - R_2(t))/(1 - R_1(t))$. By assumption this is less than 1 at $t = 1$ (the smaller *pfd* implies the lower probability of failure); when for some t it exceeds 1, it indicates that the system with the "worse" mean *pfd* is *the more reliable* over that number t of demands. If, although $q_1^* > q_2^*$, pfd_1 has the best case distribution from Thm. 2, and pfd_2 the worst case distribution from Thm. 1, this ratio is

  $$\frac{1 - R_2(t)}{1 - R_1(t)} = \frac{1 - (1 - q_2^*)^t}{q_1^*} \tag{3}$$

  which, while less than 1 at $t = 1$, will exceed 1 (will show a "reversal of roles") for sufficiently large t, and *keep increasing* with increasing t. In fact, we can Taylor-expand the numerator here, to first and second order in q_2^*, which two approximations are known to be upper and lower bounds. This yields bounds

  $$\frac{tq_2^*}{q_1^*}\left(1 - \frac{t-1}{2} q_2^*\right) < \frac{1 - R_2(t)}{1 - R_1(t)} < \frac{tq_2^*}{q_1^*}\,. \tag{4}$$

  As t increases, and while tq_2^* remains small, tq_2^*/q_1^* is a good approximation to the size of this most extreme case ratio of failure probabilities (over t demands) of the two systems, which increases linearly with t: the "reversal of roles" will occur when t exceeds approximately q_1^*/q_2^*. Further, in general, irrespective of the size of tq_2^*, there exist pairs of *pfd* distributions that increase this ratio to at least the size of the left hand bound in (4).

4 Corollaries and extensions

4.1 Reliability in continuous time

If we consider a system with exponential reliability $e^{-\lambda t}$ conditional on a failure rate λ, and the latter is seen as a r.v. Λ with probability density function $f_\Lambda(x)$ and mean λ^*, the lowest reliability compatible with these assumptions is still obtained by assuming that λ is deterministically equal to λ^*:

$$R(t) = \int_0^\infty e^{-\lambda t} f_\Lambda(\lambda)\, d\lambda \geq e^{-\lambda^* t} = e^{-\int_0^\infty x\, f_\Lambda(x)\, dx\, t} \tag{5}$$

The proof (Ap. A) is practically identical to that of Thm. 1.

Indeed, this pessimistic bound holds for any reliability function (in discrete or continuous time) that is *convex* w.r.t. a single parameter q.

As for upper bounds on reliability, the tightest upper bound under the constraint $\mathrm{E}(\Lambda)=\lambda$ is 1, which is not useful (see Ap. A, Thm. 3).

In the rest of this paper, we will keep referring to the case of an on-demand system, with its *pfd* and its probability of failure over t future demands; but our considerations will generally apply to the continuous-time case as well.

4.2 Inference from failure-free operation or acceptance tests

In the case that the system $\mathcal{S}$ is subjected to a certain number d of statistically representative, independently selected demands and exhibits no failures (the case of greatest interest for critical systems), the posterior distribution for $\mathcal{S}$'s *pfd* is:

$$f_Q(q \,|\, d) = \frac{f_Q(q)(1 - q)^d}{\int_0^1 (1 - x)^d f_Q(x)\, dx} \tag{6}$$

and Pr(surv. $\geq t$ further demands | surv. d test demands)

$$= \int_0^1 (1-x)^t f_Q(x \mid d)\, dx = \frac{\int_0^1 (1-x)^{t+d} f_Q(x)\, dx}{\int_0^1 (1-x)^d f_Q(x)\, dx} \tag{7}$$

A prior distribution that is "more pessimistic" than another may not necessarily lead to a more pessimistic posterior distribution after observing no failures. However, this does occur with the two "extreme" distributions in Thms. 2 and 1:

1. with the "most optimistic" prior distribution of Thm. 2, one test without failure is sufficient to infer that S's *pfd* is 0 with probability 1, so that the reliability function is identically 1;

2. the "most pessimistic" prior from Thm. 1, in which S's *pfd* is q^* with probability 1, will not change no matter what behaviour is observed in testing. *Any* other prior with the same mean *pfd*, on the other hand, would change as a consequence of observing 0 failures, and would change in such a way as to reduce its mean *pfd* to some new value $q' < q^*$. Therefore, it would lead to a reliability function no lower than $(1-q')^t$, in its turn greater than $(1-q^*)^t$.

These two bounds on the reliability after successful testing are somewhat uninteresting, since the upper bound is trivially 1 and the lower bound is obtained by ignoring any number of successful tests. Still, it is useful to observe that these are the only constraints on the posterior reliability that follow from stating a certain prior mean *pfd*. Using the mean alone to characterise a prior distribution is conservative, which may be appealing, but also makes it impossible for positive evidence to improve reliability predictions (while certainly avoiding errors in the direction of optimism).

We discuss in Section 5.3 more general cases of "broader" priors leading to better posterior reliability.

4.3 Effects of acceptance testing on improving systems

If there is a possibility of testing a system further after choosing it and before actual operation, the "best" and "worst" distributions of section 3 retain their respective roles. We have seen above the effects of inference about the *pfd* if there are no failures. As for the case in which failures do occur, this must be treated differently depending on what $\mathcal{S}$ is. We consider two possibilities: $\mathcal{S}$ is an off-the-shelf, replaceable physical item; or it is a design, e.g. software, subject to design-caused failures only. In the former case, if $\mathcal{S}$ fails on test, it will often be discarded and replaced with a nominally identical item, i.e., one from the same distribution of *pfd* values. In the latter case, $\mathcal{S}$ will be fixed to eliminate the "bug" that caused the failure.

Physical items When an item is chosen at random from a stock for use, and tested prior to deployment, failures are most likely, as we have seen, with the "worst-case" distribution in which the system $\mathcal{S}$ has *pfd* equal to q^* with probability 1. Then, if a failure does occur, and $\mathcal{S}$ is a physical item, it will be substituted with another, nominally identical item: no improvement follows the failure[5]. Given the best-case distribution, instead, the substitute item will still have 0 *pfd* with probability $1-q^*$, will be tested, and rejected if its *pfd* is not 0. Eventually, a "perfect", 0-*pfd* $\mathcal{S}$ will be accepted into service, although testing and discarding imperfect ones may take some time (geometrically distributed). Even excluding the possibility of the extreme value *pfd* $= 0$, a broad *pfd* distribution will give the advantage, over a narrower, "more predictable" distribution, that acceptance testing will make the distribution of the *pfd* of $\mathcal{S}$ items actually accepted into service better than that of $\mathcal{S}$ items as produced. In other words, a broad distribution means that the population of $\mathcal{S}$ items includes with non-negligible frequency both items that are "substantially

[5] We show here the implications of this policy under the assumption that the selection of the replacement item is independent of that of the item that failed. More complex models could account for correlations, e.g. due to effects of deterioration, if the two items were stored together and different warehouses may be exposed to different causes of deterioration.

better" and items that are "substantially worse" than average. Acceptance tests will tend to select (statistically) the better items. The broader the distribution, the farther apart the worst and best items will be, and the easier it will be for acceptance testing to weed out the worse items, so that the items that pass the acceptance tests will be better – on average.

In detail, the *pfd* distribution for an $\mathcal{S}$ item that eventually passes acceptance testing can be obtained by considering that each new item is selected independently of any other item that may have failed acceptance tests. Since we already know the distribution for the *pfd* variation from which we have selected, the fact that other independently selected items failed the acceptance test gives us no information about the next item selected. In contrast, the fact that the latter passed the acceptance test does give us some information: this *pfd* distribution *after* passing d demands in acceptance testing will be the posterior distribution indicated before in section 4.2:

$$f_Q(q \,|\, d) = \frac{f_Q(q)(1-q)^d}{\int_0^1 (1-x)^d f_Q(x)\,dx}$$

So best and worst-case priors are again as in Thms. 1 and 2.

Software or design items If $\mathcal{S}$ is a design or software, the consequence of an acceptance test failure is repair, and thus a new *pfd* distribution that depends on our assumptions about the design fixing process: usually that repair will more likely improve design than make it worse (otherwise we would not attempt the repair). Then, the kind of reliability growth to be expected would depend not only on the *pfd* of $\mathcal{S}$, but on how many faults contribute to it, and the amounts of their individual contribution. A discussion under simplifying assumptions is in [6].

4.4 Resumption of operation after failure: predicting the number of failures

The results of section 3 apply to the scenario of primary interest to us: one in which the first failure in operation is catastrophic (or very expensive) so that it will end the operational life of the item that failed.

By way of comparison, we consider briefly a scenario in which one is interested in the number m of failures (as a mean, or as a distribution) over a certain number t of demands. The distribution of this number, conditional on the value of the *pfd*, is binomial:

$$\Pr(m \;\; \text{failures}) = \binom{t}{m} \int_0^1 q^m (1-q)^{t-m} f_Q(q)\,dq$$

The mean number of failures, conditional on the *pfd*, is (according to a standard result about the binomial distribution):

$$E(\text{number of failures in } t \text{ demands} \,|\, \mathit{pfd} = q) = tq\,, \qquad \text{and hence}$$

$$E(\text{number of failures in } t \text{ demands}) = \int_0^1 tq f_Q(q)\,dq = tq^*\,.$$

Substituting the mean q^* in the expression for $\Pr(m \text{ failures})$ does not offer a general upper or lower bound for this probability. We cannot generalise our Jensen-based bounding argument to bound probability of observing m failures during t demands, because the above binomial probability of m failures conditional on q is not, for $1 \leq m \leq n{-}1$, either a convex or a concave function of q throughout the whole interval $q \in [0,1]$. The same is true of the *cumulative* binomial distribution $\Pr(\leq m \text{ failures} \,|\, \mathit{pfd} = q)$.

5 When does a "broader" distribution of the *pfd* imply higher reliability?

Thm. 1 shows that the *most concentrated* possible *pfd* distribution compatible with a given mean – the distribution where the *pfd* equals its mean with probability 1 – yields a global lower (i.e., pessimistic) bound on long term (lifetime, mission) survival probability.

This suggests that in some way a "broader" probability distribution of *pfd*, for a given average, implies better survival probability. Is this true in some precise, general sense?

An issue is that there is no general meaning of "broader". Which of two distributions is "broader" is often in the eye of the beholder. Variance is often used as a measure of distribution "broadness"; but lower variance does not guarantee lower reliability (except in the extreme case of zero variance): we can produce two distributions with the same mean and variance, and yet leading to different reliability functions.[6] However, we now show that there is a specific, *non-parametric* meaning of "one distribution being broader than another" that *always* implies greater reliability, irrespective of membership or not of parametric distribution families, provided the two distributions are comparable in this special sense. Indeed, we observe that changing the probability density function for a *pfd* by *shifting probability masses apart* from each other, towards opposite extremes of the range[7], while keeping the mean *pfd* constant (Figure 5), creates a new probability density function that: (i) is "broader" in the common sense of the word, and (ii) also implies higher reliability, for any $t>1$.

This result is mathematically interesting; its practical use is to confirm that apparent differences in "broadness", in the intuitive sense, are useful cues for situations in which reasoning on the basis of the mean may be misleading. Whether the situation is indeed so can be verified by just computing the reliability function using the complete distribution of the *pfd*, if known. If unknown, then the result presents an opportunity to bound system reliability by arguing that some stated *pfd* distribution (not necessarily one of our two extremes of Thms 1 & 2) constitutes a "*pfd*-distribution bound" in the sense of our "broadness" partial ordering. The "moving probability masses apart" operation that produces this partial ordering is illustrated below, while the complete proofs are provided in Ap. B.

5.1 An operation that "moves probability mass apart" and increases system reliability, while preserving E(pfd)

For notational concision, we work in this section with the variable $U = 1 - Q$ rather than Q, the *pfd* itself; and let p denote the probability density for U. Fig. 5 represents one possible such probability density function, p, and another, p_1, obtained from p by the "broadening" operation. The two probability density functions p and p_1 only differ over some subsets of the abscissa (u) axis. Where they differ, the original density p is shown with a grey line and p_1 with a darker line.

We define the operation of "broadening" this distribution p, in intuitive terms, as **taking two separate "chunks" of probability and moving them apart from one another** (without changing their 'shapes'[8]) by displacements inversely proportional to their associated probability masses (so the lighter mass is moved proportionately farther).

[6] In *some* common cases, lower variance *does* consistently imply lower reliability. E.g., for the 2-parameter Beta distributions, increasing *pfd*-variance while keeping *pfd*-mean constant will always increase system reliability for a given number of demands $t>1$. In the case of failure rate λ in continuous time, the same is true of the Gamma family of distributions for λ. The broadening operation we introduce in section 5.1 *does* always increase the *pfd*-distribution variance. It holds constant the first moment $E(pfd)$, and its increase of variance is an instance (the t=2 instance) of inequality (16): because the increased reliability is equivalent to increases of all of the non-central moments $E(pfd^t)$, for $t = 2, 3, \ldots$, and hence, in particular, of $E(pfd^2)$, therefore of the variance.

[7] The two probability masses may be associated with any two values (or intervals of values) of *pfd* before the "shifting apart", provided the two do not overlap (*cf* Fig. 5). They might even be in one tail of the distribution, so that then one mass will be moved further into this tail, and the other mass will be moved toward the centre of the distribution. This may not intuitively look like "broadening" the distribution, but the reliability increase, proven as Thm. 4 below, still follows.

[8] This restriction is actually unnecessary, but it simplifies the definition of the operation and the proof.

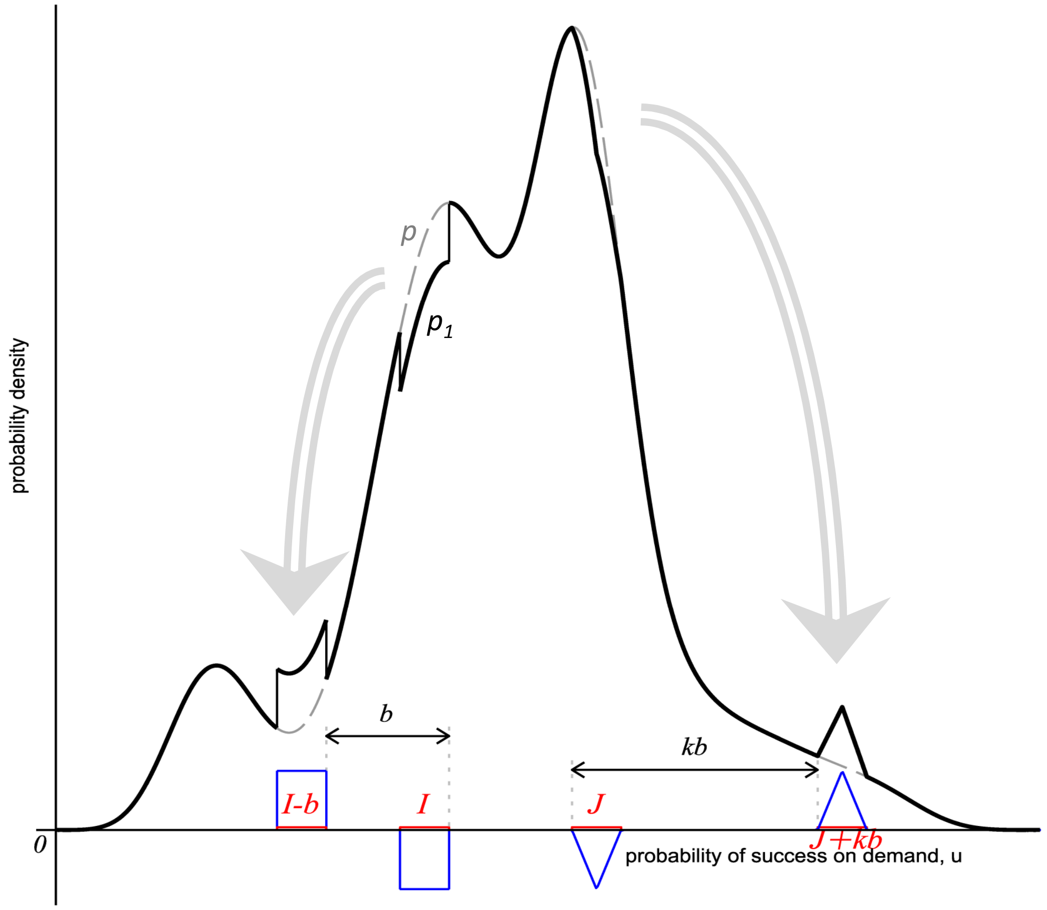


Fig. 5: "Moving mass apart" illustrated as an operation in terms of density functions. The "chunks of probability mass" being moved are a rectangular chunk at abscissae $u \in I$, which is moved to the left, and a triangular chunk at abscissae $u \in J$, which is moved to the right. The area of the former is k times the area of the latter, and thus the latter must be moved a distance k times greater, for the mean value of U to remain unchanged.

The movement of probability mass producing p_1 from p "broadens the prior distribution" of U whilst maintaining the mean.

A formal, more detailed description is in Ap. B, where the following theorem is proved:

THEOREM 4. *If a distribution p of the per demand probability of success U is "broadened" according to the procedure illustrated in Fig. 5 that leaves the mean* pfd *unchanged, then the resulting reliability over a lifetime (or mission) of length $t > 1$ demands, always* increases. *A polynomial form of a lower bound on the size of this reliability increase may be obtained, and is shown as equation (16).*

This discussion is in terms of continuous probability density functions. The theorem is also valid if the probability distribution includes (or is wholly made of) discrete probability masses (Dirac's deltas).

5.2 Partial Ordering among distributions

We observe that applying the "broadening" operation repeatedly, starting with the degenerate distribution in which $pfd = q^*$ with certainty, defines a totally ordered family (or "chain" in the terminology of mathematical order theory) of distributions with monotonically increasing values of reliability (for any given future time).

Furthermore, starting from any distribution p that is not the "best case" distribution of Thm. 2, there are multiple degrees of freedom in choosing how to "broaden" it according to the above definition of "broadening". Thus, there is an infinite set of such totally ordered chains of distributions. If two distributions are in the same chain, that is, one can be obtained from the other by a finite sequence of applications of the "broadening" operation[9], then Thm. 4 tells us that the broader distribution implies higher reliability.

Given the continuity of the reliability functions we consider, one can also see that "broadening" can under very general conditions "compensate" for a higher expected *pfd* so that, between two distributions with different mean *pfd*, the one with higher (worse) mean yields, for sufficiently large t, better reliability than the one with the lower (better) mean, as in Figure 2. This reinforces the message of section 3.1 about how little mean *pfd*s imply about reliability.

5.3 Posterior reliability after observing failure-free operation

We observed in Section 4.2 that the two extreme priors with the same mean, producing the lowest and the highest reliability functions, also lead to the lowest and the highest posterior reliability functions after observing failure-free operation. One can easily construct many examples in which the same ordering applies to the posterior reliability between a "broader" and a "narrower" distributions produced as in section 5.1: it is often the case that inference based on the "broader" prior distribution produces a posterior distribution that implies better reliability than that obtained by performing the same reasoning, given the same amount of failure-free operation, on a "narrower" prior distribution with the same mean. Therefore, given the continuity of all the functions considered, it is also possible to find examples in which a prior that has worse *pfd* than, but is broader than, another prior, leads to better posterior reliability.

That "broadening" the prior distribution of the *pfd* improves posterior reliability given failure-free operation is not a general truth however: one can build counterexamples. The important conclusion is that since such an effect *is possible*, and calculating posterior reliability is easy, one should not assume *a priori* that the ordering between the prior mean *pfd*s of different systems implies the same ordering for posterior reliability, but calculate them for each individual case.

6 Generalisations and limits: multiple components or parameters

The above results apply to systems for which a single, "black-box" reliability parameter is sufficient description of the failure process and of the uncertainties about it. How useful are they when the system under consideration is described by *multiple reliability parameters*, e.g. a multi-component system in which each component adds one or more parameters to a model of the system's reliability?

For brevity, we only discuss generalisability for the pessimistic bound on reliability of Thm. 1 (and § 4.1): if we take the expected values of all the reliability parameters, do we always obtain a pessimistic bound for system reliability?

The answer turns out to be "No". The bound applies in various restricted scenarios (we give examples in the next subsection), but not in general as in the case of a single parameter, and in 6.2 we show that the exceptions include very basic system configurations.

6.1 Sufficient conditions for the pessimistic bound to apply

As a premise, we note that this pessimistic bound on reliability applies because the reliability function of interest is a *convex function* of the parameter throughout the latter's range of possible

[9] In standard mathematical terminology the two distributions are then *comparable* in the *partial order* which is the *transitive closure* of our broadening operation, viewed as a binary relation among probability distributions on $[0, 1]$

values (Ap. A). The concept of "convex function" extends to functions of multiple parameters [14, Ch. 3], [10, Ch. 3]; we will use some implications of this generalised convexity property.

We consider a system with multiple components, each contributing a parameter q_i to the expression of the system reliability, $R_{sys}(t; q_1, q_2, \ldots, q_n)$, and give examples of *sufficient* conditions under which the bound does apply, that is, using the expected values of the uncertain parameters as if they were the true values yields a pessimistic bound on system reliability. Cases in which the bound applies include:

1. if (i) *uncertainty only affects one parameter* (one component), (ii) the reliability of that component is a convex function of that parameter (as in our example of a geometric reliability function, determined by the *pfd*), and (iii) the reliability of the system is a *monotonically non-decreasing function* of the reliability of that component. For instance, this covers the case of uncertainty about a single component of a non-repairable series, or parallel, or k-out-of-n system, made of independently failing components. The bound applies simply because monotonicity implies that the lowest possible value of the argument gives the lowest possible value of the function;

2. if (i) *uncertainty affects multiple parameters*, but (ii) *independently* (i.e. the parameter values are independent random variables; i.e., the joint uncertainty about them all is without 'epistemic correlation'), and (iii) the true system reliability is a product of convex functions, each of one only of these unknown parameters. For example, the reliability of a multi-component, non-repairable *series* system in which all components fail independently factorizes as $R_{sys}(t; q_1, q_2, \ldots, q_n) = \prod_{i=1}^{n}(1 - q_i)^t$, in terms of the n unknown component *pfds* q_i. The pessimistic bound applies here because independence implies that we can obtain $R_{sys}(t)$, the mean (expected value) of this product, as the product of the means of its n factors. To each of these our previous argument applies; the most pessimistic distribution of each parameter q_i is the one concentrated at its mean $E(q_i)$. Each factor in the R_{sys} product is thus minimised, and so is, therefore, the product R_{sys} itself.

3. if (i) *uncertainty affects one or more parameters*, and (ii) system reliability is a *convex* function $R_{sys}(t; q_1, q_2, \ldots, q_n)$ of *the vector* of these parameters, then again the bound applies. The bound applies as a multivariate case of Jensen's inequality, by direct analogy with Thm. 1. As a particular instance of this case we conclude also…

4. if (i) *uncertainty affects one or more parameters*, (ii) system reliability is $R_{sys}(t) = (1 - q_{sys})^t$ (or any other convex, non-decreasing function of the system *pfd*, q_{sys}), (iii) the system *pfd* is a function of those parameters affected by uncertainty, say $q_{sys} = S(q_1, q_2, \ldots, q_n)$, and (iv) *this latter function S is a concave* function of its *vector argument*, then again the bound applies: using the expected values of all parameters yields a pessimistic bound on reliability. The bound applies as an instance of case 3 above, because any nondecreasing convex function of a scalar value, that is itself a convex function – here $(1-S(q_1, q_2, \ldots, q_n))$ – of a vector of other variables, is itself a convex function of the latter, the condition for case 3.

General notes (i) all the sufficient conditions listed need to be true just for those values of t for which one wishes to estimate a bound on reliability – although in many scenarios, including those of Thm. 1 and equation (5), they hold for any $t \geq 1$; (ii) similar sufficient conditions for the pessimistic bound to apply exist for the case of *continuous*-time reliability: they can be stated by substituting "failure rate" for *pfd* in the statements above. For concision we omit the detail, which uses closely similar reasoning.

The differences between these cases are subtle, but highlight possible surprises when attempting to generalize the pessimistic bound. It might seem natural that if a system's *pfd* is a *monotonic* non-decreasing function of the *pfds* of the components – the very common situation of an architecture in which "*improving any component will improve the system*" – then the pessimistic bound would extend. But case 3 above indicates that the sufficient condition for our bound to hold is *convexity* of the system *pfd* as a function of the vector of component *pfds*, rather than *monotonicity* as a function of each component *pfd*. If the sufficient conditions above seem inconveniently restrictive, one has to consider that the result they guarantee is very strong: it is a pessimistic bound that holds given *any joint probability distribution of the parameters*, including some cases of 3 and 4 for which intuition is of little help, e.g. when some parameters are highly correlated, and/or some parameters are negatively correlated. In case 2 a special joint distribution – epistemic independence – plays a role. Epistemic independence (learning about the true *pfd* of a component does not affect one's beliefs about the others) is reasonable to assume in some cases, e.g. physical components randomly sampled from the output of a production line with known distribution of component *pfd*, but not in many others.

We next show examples which underscore that these results cannot be assumed to remain true under even slightly different assumptions.

6.2 Examples where the pessimistic bound does not apply: series and parallel systems with independent failures

It is tempting to assume that the sufficient conditions for the pessimistic bound to hold extend easily to similar cases, e.g. that the statement of case 1 about series, parallel, etc systems with uncertainty about only one parameter extends to multiple parameters, possibly using the sufficient condition of case 3 above.

But we do not in general find the functions that give system reliability for these simple redundant configurations, in terms of component reliability parameters, to be convex or concave over the domains of interest. That is, *given the architecture of the system but not the detailed joint probability distribution of the parameters, the pessimistic bound may or may not apply.* For any reader who may be interested, the remainder of this section exhibits two simple multi-parameter cases, allowing general forms of epistemic correlation of the parameter uncertainties, which can cause the pessimistic bound not to apply.

6.2.1 Series system with independently failing components

We return to case 2 above of series system with independently failing components, but remove the condition of epistemic independence in the parameter uncertainty. To simplify the problem, we limit ourselves to a series system made up of just *two* components A and B that fail independently: given specific *pfd* values of the two components q_A and q_B, the system's *pfd*, q_{AB}, is the product $1 - q_{AB} = (1-q_A)(1-q_B)$ and its probability of survival for t demands is

$$\Pr(\text{survival for } \geq t \text{ demands} \,|\, q_A, q_b) = (1-q_A)^t(1-q_B)^t \tag{8}$$

and thus when averaging over the uncertainty it is:

$$\Pr(\text{surv. } \geq t \text{ dem.}) = \int_0^1\int_0^1 (1-q_A)^t(1-q_B)^t f_{Q_A,Q_B}(q_A,q_B)\, dq_A dq_B$$

Were the parameterised reliability function (8) convex over the unit square, then the 2-dimensional version of Jensen's inequality would be applicable. However, we show in Ap. C that this is not the case:

THEOREM 5. *For no t is the function defined in equation (8) either convex or concave on any open subset of the unit square.*

It follows that, for each $t = 2, 3, \ldots$, there exist bivariate distributions of (Q_A, Q_B) such that Pr(survival for at least t demands) exceeds $(1-q_A^*)(1-q_B^*)$, and, for that same t, there exist other distributions for which this inequality is reversed.

6.2.2 "Parallel" system with independently failing components

In the case of two components in *"parallel"* (and still assuming failure independence), the situation is similar, though with a slight complication. For such a system, system *pfd* is $q_{AB} = q_A q_B$, which when averaged over our uncertainty gives

$$\Pr(\text{survives} \geq t \text{ demands}) = \int_0^1 \int_0^1 (1 - q_A q_B)^t f_{Q_A,Q_B}(q_A, q_B)\, dq_A dq_B$$

We show in Ap. C that:

THEOREM 6. *The system reliability function* $(1 - q_A q_B)^t$ *is, for any fixed* $t \geq 2$*, convex on a convex subset* F *of the unit square precisely if* F *is a subset of the region* D *on which this reliability* $(1 - q_A q_B)^t$ *has values below* $(1 - \frac{1}{2t-1})^t$*.*

This theorem identifies conditions for the bound to apply, but not very useful ones. Precisely, the bound

$$(1 - q_A^* q_B^*)^t \leq \int_0^1 \int_0^1 (1 - x)^t (1 - y)^t f_{Q_A,Q_B}(x, y)\, dxdy$$

only applies if the system reliability has probability 0 of exceeding a value that depends on t but is lower than approximately 0.6. Hence, in the parallel case, it is not that the worst case bound based on expected values is never applicable; but rather that it can be applicable only for very unreliable systems, and it is thus irrelevant for most applications.

6.2.3 Other multi-parameter reliability functions

Other cases with two or more uncertain parameters include single-component systems with reliability functions from a family with two or more parameters, e.g., Weibull. Again, we found that using the expected values of the parameters does not yield a general lower bound on reliability.

7 Discussion. Implications for decision making

The immediate implications of the theorems shown are:

- the reliability function for a given system can be bounded in a very simple way, obviating the problem of not knowing the *pfd* distribution in detail. Especially the pessimistic bound obtained by using the mean *pfd* is very useful when needing to avoid errors in the direction of optimism; although these bounds will sometimes be too slack for convenience;

- when evaluating a system with given mean *pfd* (or failure rate), if we define a "bounding *pfd* distribution" for the reliability parameter, according to the "broader distribution" relation and associated partial ordering of Section 5.2, we obtain a numerical bound on the reliability function (a lower bound on reliability for a bounding distribution that is *narrower* than the true distribution, an upper bound for one that is *broader*);

- when comparing two systems with the same mean *pfd* (or failure rate), if one has broader distribution in the sense of Section 5.2, then it is the more reliable system.

We discuss some further implications of these facts for various decision tasks in safety and reliability engineering.

7.1 Implication of the "mean implies pessimism" result for system assessment

In many practical decision problems in reliability and safety, guaranteeing that any error is "on the side of pessimism" is considered good practice, despite it not being so in basic decision theory. Thus this simple result is useful in many cases.

Statistics from heterogeneous populations, reliability databases If the mean reliability parameter is derived from failure statistics over a population, so that the mean hides the potential for variability of this parameter across the population, then Thm. 1 ensures that using this mean value leads to a pessimistic prediction.

Results of conceptual modelling for design reliability There are examples of useful probabilistic modelling that yield statements about the mean value of the reliability parameter obtained. For instance, the body of work (to which we and colleagues have contributed) about the efficacy of diversity against common mode failures yields useful insight about the mean *pfd* that various approaches to diversity can achieve (see e.g. [32, 39]). Likewise the literature on the efficacy of software testing often compares expected achieved *pfd* [18]. The paper [7] demonstrates ways of turning limited statements about an assessor's subjective distribution for a reliability parameter into a pessimistic expected value for the parameter. In all these cases, thanks to Thm. 1, results about the mean value of the reliability parameter imply results about *worst-case* reliability.

7.2 Expert elicitation and calibration

When depending on expert opinion as input to risk assessment, a common concern is poor calibration: the fact that experts' confidence statements about their estimates of parameters may be incorrect.

The common concern is *over*confidence: stating narrower subjective distributions than warranted by the experts' actual expertise (see e.g. [28] for recent results and references). Overconfidence is seen as dangerous: it may hide uncertainties that ought to be considered. So, much thought is given to how to avoid it. Thm. 4 identifies situations in which overconfidence is actually conservative, while one should be concerned about *under*confidence – an expert who, through modesty or because effective measures were used during elicitation to avoid overconfidence, overstated his uncertainty about a parameter – producing over-optimistic predictions.

When we have reasonable confidence in the accuracy of the expert's opinion about the mean value of the parameter, but less so about its distribution, accepting overconfident statements – distributions that are probably too narrow – can be the prudent course. As an example, one can envision a negotiation, for instance between a proponent of a safety-critical system and a regulator asked to approve operation of the system. Suppose that the parties have reached agreement about a claim about a mean system *pfd*. There remains to decide the degree of uncertainty about the true value. At this point, the proponent could stipulate, or the regulator require, so as not to make the whole case for the system rely on a possibly over-confident point estimate, that the claimed two-sided confidence intervals around the mean be wide, so that any error is in the "conservative" direction. This constraint, imposed in the name of prudence, will actually cause more optimistic reliability predictions than using the mean *pfd*.

7.3 A decision scenario: component selection with regard to probability of physical failure

Suppose we have to select a hardware system $\mathcal{S}$ from two off-the-shelf choices, with identical nominal, empirically demonstrated *pfd*, i.e., that have given the same number of failures per year of operation, over very extensive operation: type $\mathcal{S}_1$, about which we have evidence of rather narrow *pfd* distribution, ensuring good predictability of behaviour, and type $\mathcal{S}_2$, which despite having exhibited the same average *pfd* as $\mathcal{S}_1$, is known for evidence of somewhat erratic quality, with some items exhibiting early failures and others extremely high reliability.

How we should react to this information depends, of course, on our priorities. For instance:

- if we were only concerned with predicting the total number of failures over a long period of usage of a large population, with use continuing after each failure, we would be indifferent between types $\mathcal{S}_1$ and $\mathcal{S}_2$: for both the total number of failures will be narrowly distributed around a value determined by their common average *pfd*;

- if the problem were one of setting up stocks of spares next to each individual installation of $\mathcal{S}$ or of deciding preventive maintenance intervals, the predictability of type $\mathcal{S}_1$ might be preferred;
- but in our reference scenario, system $\mathcal{S}$ is a critical system, an individual instance of which must have a very low probability of failure as preventive maintenance is impractical and failures are likely to be catastrophic (and will end the system's life). Here, counter-intuitively, the erratic quality of type $\mathcal{S}_2$ is an *advantage*: it means a lower probability of observing a catastrophic failure over the lifetime of each instance. In other words, between type $\mathcal{S}_2$ and type $\mathcal{S}_1$, type $\mathcal{S}_2$ gives the higher probability that the individual system actually installed will be so reliable as not to fail at all over the installation's lifetime; it also gives higher probability that an individual system will be so unreliable that it would fail multiple times, if use restarted after each failure[10], but this latter scenario has no importance since any failure would end the life of the system.

We are using here the informal terms "narrow" and "broad" for distributions of the *pfd*. More generally, even if $\mathcal{S}_1$ and $\mathcal{S}_2$ have different mean *pfd*s, "broader" distributions may imply advantages in terms of reliability. Section 5 defines a meaning of "broader" that guarantees this effect. When the assumptions of Thm. 4 do not apply, a distribution that just *looked* "broader" than another one with the same mean may not imply higher reliability, or, instead, a distribution with worse (higher) mean could imply higher reliability thanks to being broader, as in the extreme example at the end of section 3. The message here is simply that when dealing with *pfd* distributions that exhibit very different shapes, even outside the precisely defined scenario of Thm. 4, one should not just compare their mean or some percentile, but calculate their actual effects on reliability.

7.4 Decision scenarios: systematic failure, software

If we are dealing with the *pfd* due to systematic failures only, as e.g. with software or more in general with design faults, the distribution of *pfd* represents, for an assessor, a subjective probability distribution that is meant to summarise the epistemic uncertainty about the true *pfd*. In its turn, this subjective probability distribution may be seen as caused by the aleatory uncertainty that affects software development, so that development processes that are apparently identical may deliver products with different quality. For instance, if assessors know that past products, that had identical records of the development practices applied and of the results of the verification steps performed, nonetheless exhibited very different levels of reliability, their subjective distribution for the *pfd* of a new product with a similar record needs to be a broad one.

7.4.1 Implications for the management of software/system development

The "best" distribution in Thm. 2 is one in which most attempts to produce the software or design required would produce a "perfect", fault-free result, but a few produced software (or a design) that is guaranteed to fail. More generally, the results in section 5 imply, perplexingly, that a less predictable development process is better than a more predictable one delivering the same average. This observation has little value for most decision-makers, e.g. project managers who need to decide how to ensure high reliability with respect to design faults, because their knowledge of the *pfd* to be expected from the available combinations of methods is usually vague, even in terms of mean *pfd*. However, it casts a new light on the general issue of risk-taking in projects of a critical nature.

For instance, we can apply this observation, speculatively, to the case of the Ariane V [29], which was at first developed with a design fault that ensured 100% probability of accident at the first launch (*pfd* equal to 1, if we call a launch "a demand" and restrict "failure" to "destruction of the vehicle"). After the accident, the management of the project was widely blamed for omitting those elementary checks that would have discovered this fatal design fault. One might defend the management decisions that led to that design if it could be argued, upon further study, that any

[10] This is so because the average number of failures over a long time must be the same between $\mathcal{S}_1$ and $\mathcal{S}_2$. So, whichever has higher probability of zero failures must also have higher probability of *n or more* failures for some $n \geq 2$.

feasible alternative decision within the same budget constraints was likely to reduce the probability of that disastrous design flaw giving $pfd = 1$, but only at the cost of increased probabilities of other design flaws, all associated with lower values of system *pfd*, so that those alternative decisions would lead to the same (or worse) overall mean *pfd*, with a narrower distribution. This reasoning runs against common opinion and most expert advice.

If the first Ariane V flight had failed due to a failure mode that in retrospect had 10^{-3} failure probability per flight, criticism of project management would have been much less sharp than it was in fact after the 100% probability of failure was revealed; we could have heard comments that after all the flight was afflicted by unusual bad luck. The mathematics just shown here indicate an alternative interpretation: the presence of a high-probability failure mode rather than lower-probability ones is not *ipso facto* an indication of worse project management decisions. It might instead indicate simply that the development process exhibits high variability of results, and although we may instinctively prefer a predictable process, we have no evidence that even the methods that we consider "best practice" can guarantee predictably better mean *pfd*, and better expected reliability, than what was achieved by the process that was actually used (which includes some risk of $pfd = 1$). Then, if the decision goal is to minimise the probability of accident, a development process with greater variability of resulting *pfd* may sometimes be a better strategy than one whose outcome is very likely to be close to the mean. Unfortunately, project managers generally lack the information to decide which such risk-taking is justified.

7.4.2 Implications for system acceptance criteria

We now change our viewpoint from that of someone charged with developing a system to that of a decision maker in the process of approving a safety-critical system for operation – for instance a safety regulator. We discussed earlier (§7.2) the risk of interpreting broad confidence intervals on *pfd* as being conservative, while they actually imply optimism on reliability. We now consider instead the case in which the decision maker correctly propagates uncertainty on *pfd* to compute reliability predictions.

Suppose that the usual steps of analysis led us to believe that a certain function in a system must have a $pfd \leq 10^{-5}$. If we follow a standard like the IEC 61508 standard for functional safety [24], and try to take into account the inevitable uncertainty about the true value of *pfd*, this requirement may have different interpretations, including at least:

- a constraint on the mean *pfd*: $E(Q) \leq 10^{-5}$; or
- that there should be a minimum confidence that the *pfd* does not exceed the stated bound: $Pr(Q > 10^{-5}) \leq \alpha$.

With both interpretations, we can construct scenarios in which the choice is between two options (distributions of the *pfd*) such that both satisfy the requirement, but the option that satisfies it by a narrower margin (that is, with the broader distribution) ensures better probability of accident-free lifetime. Yet, if an accident did happen and it was found, after the fact, to have had a high probability for that specific system, those in charge of design and acceptance (and licensing if the system is subjected to a licensing regime) would risk condemnation for allowing a highly dangerous system into operation.

Thus, all the actors in the development, acceptance and, where applicable, regulation process have incentives for preferring the narrower distributions, despite this being against the apparently reasonable criterion of minimising the expected number of accidents.

7.4.3 Probability of $pfd = 0$

The "best case" distribution in Thm. 2 implies a non-zero probability $1 - q^*$ of the system having $pfd = 0$. This is implausible for hardware, but for software it simply means not containing design faults: we may call such software "correct", "perfect", or "fault-free". Still, a claim of this kind is likely to be controversial and needs some comments.

We hasten to add that the considerations in this paper do not rely on "perfection" being feasible. Even if one stipulates that the case of $pfd = 0$ must have zero probability, the "best" distribution of Thm. 2 can be seen as a limit for families of distributions which obey this constraint, and $(1 - q^*)$ remains a practical upper bound for the reliability over an arbitrarily large number of future demands: consider that for any assumed number of demands t and any arbitrarily small ϵ, one can assign the probability mass $(1 - q^*)$ to a range of q values such that $1 - (1 - q)^t < \epsilon$. Whether the distribution thus obtained could be argued in practice to be correct is unrelated to whether $pfd = 0$ is possible, and so are our more general qualitative conclusions about the advantages of certain broad distributions.

Returning to claims of non-zero probability of $pfd = 0$ for software (or design in general), however, we believe that they deserve consideration. Many think that any claim of defect-free software is unbelievable, on the basis of experience and of the usual complexity of software. On the other hand, large parts of the software engineering community (especially those studying and practicing "formal methods") point at it as a very feasible target[11].

We think that the case of a "perfect", "fault-free" design is actually plausible, especially for very simple software as required in some critical applications, and that reasoning about its probability (for the class of faults of interest, e.g. those faults that might cause accidents) is a way forward for assurance in such applications. Such "perfection" is not easy to achieve, and no known method would give 100% probability of achieving it. So crucially, a claim about $pfd = 0$ should include a probability of this being the case; and this should be based on solid arguments, invoking for instance statistical evidence about the effectiveness of development and verification methods. The arguments in favour of using this concept in safety arguments include [5]:

- in current practice and according to most standards and regulations, most of the evidence used to argue that software is unlikely enough to exhibit dangerous failures is anyway more suitable to claim $pfd = 0$ than a low, non-zero *pfd*. Indeed, this evidence concerns the methods applied to avoid design faults in construction, find them in the finished product, or tolerate them at system level. These methods are not selective between faults with a higher or lower probability of manifestation in operation. Thus, such evidence can only be relevant towards a claim of very low probability of failures if it supports a claim of the software having *no* residual faults, and thus *that very low probability being zero*;

- in addition, for long operational lives, seemingly modest probability of fault freeness will give better assurance of reliability than a seemingly more extreme claim on *pfd*, which also poses great difficulties in building a sound argument [5, 33, 34];

- if instead we assume that a *pfd* equal to 0 is impossible, rather than difficult to achieve and demonstrate, then to demonstrate low probability of dangerous failure over a long operational life[12] we would need to test for some (often impractical) multiple of that life, or to have implausibly strong prior beliefs [33, 34, 41].

7.5 Is predictability good or bad? Is confidence optimistic or pessimistic?

The theorems in sections 3 and 5 point to the non-intuitive fact that distributions that appear "broader", that is, "less homogeneous", implying "less predictable" products, may *imply better reliability*. The "worst case" distribution of Thm. 1, in which the value of the reliability parameter is certain and identical to the mean, can be seen as the "most homogeneous" or "most predictable" distribution, among those with the given mean. In sampling from a population (e.g. a population

[11] And sometimes simplistically assume that mathematical proof would remove all difficulties in trusting the software. In reality, proofs may turn out to be erroneous or the requirements that have been proved to be satisfied may turn out to be erroneous or incomplete.

[12] E.g. the cited civil aviation requirement of "not anticipated to occur during the entire operational life of all airplanes of one type", which when translated into failure rates leads to requiring upper bounds of the order of 10^{-9} per hour or per flight [16].

of nominally identical off-the-shelf hardware components), it means that every item picked will have the same *pfd*. As a subjective distribution, it is the one in which the subject has the most confidence in his estimate. Increasing the opposite tails of a distribution (a special case of our "broadening" in Thm. 4) means reducing confidence levels for confidence intervals close to the mode of the distribution.

Confidence and predictability have intuitive positive connotations. Much of the engineering mindset is about achieving predictable results, in design, manufacturing processes, etc; in software engineering, much emphasis (and investment) has gone into initiatives that attempt to limit the variability of the results of software development processes. These theorems propose an exception to the general validity of this preference.

Another consequence of equating "more uncertainty" with more risk is to assume that broader distributions are automatically more conservative than narrower ones: a fallacy, in this context. Examples in the literature suggest that this is often an implicit assumption. The necessary warning that being "conservative" from one viewpoint may instead be optimistic from other important viewpoints (like assessing the probability of failure or accident), as this paper demonstrates, is often omitted, or directly violated. For instance, in NUREG/CR-6823 *Handbook of Parameter Estimation for Probabilistic Risk Assessment* [3], passages addressing "conservatism" in estimating distributions include: "to be conservative the uncertainty distribution is often assigned a larger variance than the data alone would call for." [3, §6.1.3.1]; "the confidence interval is conservative, and so is wider" [3, §6.3.2.3.2]; and "the method [...] underestimates the uncertainty in the final answers [...]. Kass and Steffey (1989) present a correction [...] The plant-specific posterior means are unchanged, but the posterior variances are increased" [3, §8.2.4] — in the context of a Gamma distribution, for which this transformation is an instance of *reliability-increasing (optimistic)* "broadening". The tutorial [40] warns "Beware of conservatism" because "the degree of conservatism can vary from analyst to analyst [...] and "a conservative analyst injects his/her own values into the analysis, and [...] usurps the decision-makers role"; valid concerns, but perhaps a more pressing concern is that how to be conservative may not be obvious, as in the advice from the same tutorial: "start with a noninformative prior distribution (which, with its heavy tails, tends to be conservative)", which is wrong in the important scenarios we have discussed.

The immediate conclusion here is that decision makers need to be aware of the possibility of surprises, and be prepared to propagate the uncertainty on the distribution of parameters to the measures of actual interest, like reliability. Further considerations follow.

7.6 General decision making criteria in the presence of epistemic uncertainty

The discussion so far has assumed that the overriding concern of the decision maker is to minimise the lifetime probability of failure, a reasonable decision criterion when failure is expected to be catastrophic. The theorems we have presented are useful tools for decision making based on this criterion. However, as outlined in 7.4.2, other considerations will often come into play, even when the lifetime probability of failure is ostensibly the main concern.

This issue is most evident when we consider the decision to accept a system with a broad distribution of its reliability parameter (*pfd* or similar). Suppose that the system in question has both a mean *pfd* and a predicted lifetime probability of failure, $1 - R(t_{\text{life}})$ that are acceptable according to established criteria for that category of systems, say $1 - R(t_{\text{life}}) < \epsilon$; and that its value of $1 - R(t_{\text{life}})$ is lower than for an alternative system, which has a narrower *pfd* distribution; but that the system we consider has "large" epistemic uncertainty, for instance, one implying a 1% probability of the true *pfd* being such that the probability of accident is 50 times greater than the accepted target ϵ.

In the case of regulated industries, where safety decisions are intended to apply the risk acceptability criteria of a society, this kind of perplexing scenario is not necessarily addressed by the general principles usually stated.

We speculate that the reaction to such a scenario would vary between different people and communities but a certain parameter distribution is likely to elicit very different reactions in different contexts. We expect that in many situations decision makers would oppose accepting into operation the system hypothesised above for reasons like:

- they may be reluctant to allow a situation to develop in which the worst case consequence is one of discovering in hindsight that the system chosen was indeed substantially less safe than desired. This invites questions:
  - "anticipated regret" has been identified as a common pattern in everyday decision making [43], which may or may not lead, in a specific circumstance, to rational decisions in view of the decision maker's goals;
  - for an item with the potential of massively catastrophic failure consequences, e.g. accidents in a nuclear or chemical plant, a reluctance to accept a system with broad uncertainty may be linked to the fact that the risk of large accidents (e.g. ones causing many deaths) is usually considered by the public as less acceptable than that of a similar expected loss per year (e.g. number of deaths) distributed over small accidents. Some safety regulators explicitly include this aversion to large accidents among the socially mandated risk acceptability criteria they have the responsibility to apply. But the scenario here is different and more complex: the uncertainty does not concern the size of the possible disaster (nor its lifetime probability); it concerns the likelihood of choosing a course of action with a relatively high probability of causing that disaster, albeit actually outweighed by an increased likelihood of very low probability of causing it. What is society's attitude (if any) to this dilemma between highly abstract scenarios?
  - what if the consequence of failure is on a smaller scale, but with many more possible instances? For instance, we may consider a safety-critical part of an automobile, or an implantable medical device. Would high uncertainty then be more acceptable?[13]
  - should the decision maker's attitude differ between situations in which parameter uncertainty is directly due to "aleatory", "physical" uncertainty, and those in which it appears more essentially "epistemic"? For instance, in the scenario of section 7.3, there is an action of drawing an item from a stock of apparently identical items, while in the case of software assessment 7.4, the uncertainty seems more purely subjective.
- broad distributions may indicate substantial uncertainty *even about the distribution itself.* A concrete example is the choice between a system with a conventional, tried and tested design and an innovative one offering "in theory" increased reliability. Here, the narrow distribution for the conventional system may well be derived, via formal statistical inference, from extensive statistics in representative situations of use; the broader distribution may well represent just the fact that no assessors believe they can confidently exclude "surprises" from the innovative system, but the details of the distribution are not supported by convincing arguments. This kind of uncertainty may be important, but difficult to represent formally: how to treat it correctly is an open problem;
- last but not least, given a broad distribution, means are sometimes available for substantially narrowing it (thus reducing "anticipated regret") while improving predicted reliability, at acceptable cost. For instance, operationally realistic testing is sometimes inexpensive, and will either produce failures or a rapid improvement in the subjective distribution of the reliability parameter, and thus the predicted reliability. However, such techniques are not available for all kinds of system.

[13] With such mass produced, individual person-oriented items, there would be an additional reason for accepting high uncertainty: even if the product turned out to be the "unlucky draw" – to have high *pfd* – any failure or accident (with at most a small number of injuries or deaths) would likely lead to correcting the design factors that caused the accident, and thus to further improving the expected number of accidents over the lifetime of the item. This would be costly for the vendor, but might on average yield better public safety than a product with lower uncertainty about its *pfd*. This part of the scenario can be incorporated in the probabilistic model of the total number of accidents over the design's lifetime [8]. Yet the idea that customers of a certain product may receive a dangerous item "by lottery" could appear morally repugnant.

7.7 Decision making heuristics

Until now, we have been concerned with the correct decision in a certain scenario; we now switch to considering the likelihood that people in those scenarios would decide correctly without use of formal probabilistic reasoning.

In informal discussions we found that several people (lay people, reliability experts and psychologists) found the results of sections 3 and 5 surprising, and their implications for decisions, in some scenarios, intuitively "wrong".

It is well known that in conditions of uncertainty people may reach different decisions depending on whether they apply formal mathematical reasoning or what we may call "intuitive" judgement – well developed, not fully conscious, quick decision mechanisms, called "heuristics" in the psychological literature [27, 26]. These heuristics are well suited for satisfactory decisions in many situations, but there are well documented categories of situations in which they lead to decisions that appear inconsistent, or systematically violate the criterion of maximising expected utility. For professional decision makers, the importance of this kind of research is in helping to discriminate between situations in which "instinct" will produce the right decision and situations in which a formalised decision process is necessary. Research shows that although in some areas some experts develop abilities to perform intuitively complex probabilistic decision tasks with great accuracy, other categories of experts commit serious mistakes when not applying formal probabilistic reasoning, and *fail to recognise situations in which they need to apply it.*

We ran experiments in which lay people were presented with decision problems mimicking the choice between two systems, described in section 2. For instance, they were asked to choose between playing different games of chance, or taking different medicines with different risk of side effects. The common factors were a Bernoulli trials scenario, and a need to choose between alternate distributions of the failure probability per trial. We found that participants reacted correctly to information about the mean probability of failure per trial (alike to our mean *pfd*), that is, they preferred options in which it was lower; but reacted incorrectly to the spread in this distribution, often choosing narrower distributions with lower probabilities of winning [2].[14]

In conclusion, decision makers need to consider scenarios involving broad probability distributions as ones that require explicit propagation of uncertainty rather than summary, "intuitive" judgement, especially in view of the pervasive preference for predictability in engineering culture.

8 Conclusions

We have proved useful results for the many situations in which there is reasonable confidence about the expected value of a reliability parameter (probability of failure per demand or failure rate), but not about its distribution:

- from the viewpoint of predicting probability of failure-free operation over multiple demands, i.e. predicting a system's reliability function, using the mean of the parameter instead of its full distribution is guaranteed to err on the side of pessimism;
- there is also a bound on the side of optimism: for on-demand systems, assuming that the system has a probability q^* of having *pfd* equal to 1, and probability $1 - q^*$ of having *pfd* equal to 0;
- with these two bounds, one can both perform pessimistic calculations and have an idea of the size of the approximation error;

[14] Our decision scenarios are reminiscent of the well documented phenomenon known as "ambiguity aversion", which can be stated as: "When given a choice between two otherwise equivalent options – one in which the probability information is stated and another in which it is missing – most people avoid the option with missing probability information" [38]. "Ambiguity aversion" has been widely studied; researchers have linked to it many experimental results demonstrating apparent inconsistency in decision making; as a social phenomenon, it may explain apparent excessive purchasing of insurance; as a scientific topic, there is interest in explaining it in terms of which mental processes cause it and what may have determined its evolution as a general human behaviour. The main difference between the "ambiguity aversion" scenarios and our experimental scenarios is that in ours the probabilities of all outcomes are known in advance (or, rather, can be calculated with moderate effort).

- the two distributions of *pfd* values that produce these pessimistic and optimistic bounds also guarantee pessimism, and optimism respectively, of inferences from observing some failure-free operation or testing. This invariance does not hold for most other distributions.

These results have direct use in many situations in which there is epistemic uncertainty and one is not prepared to go through sophisticated mathematical treatment of it, either because of its cost of because the complex resulting argument is not considered helpful by the decision makers, and especially in situations in which "conservative" decisions are considered adequate.

We have also identified a partial ordering between distributions of the reliability parameter that share a common mean. This observation leads to several conclusions that may be surprising:

- contrary to the common aversion to accepting a wide range of uncertainty, sometimes a broader distribution offers better reliability;
- in particular, between two distributions with different means, the one with the worse (i.e., higher) mean may yet be the better one if it is substantially wider than the other: a system with worse expected reliability in the short term, *and wide uncertainty about it*, then offers better long term reliability than one with a better *and more predictable* short-term reliability;
- making a manufacturing process more predictable in terms of achieved short-term product reliability may damage the operational record to be expected, unless the mean of the distribution is also improved substantially;
- given a subjective distribution of *pfd* or failure rate that one does not fully trust (or that leads to excessive computational complexity) one may be tempted to "err on the side of prudence" by assuming a broader distribution, as in our section 5. In fact, this purported pessimism produces *optimistic* predictions:
  - seeking "pessimism" in the distribution of a parameter is pursuing a mirage: what matters is whether *the specific predictions of interest* err on the side of pessimism;
  - for predicting the probability of no failures over a time span, "narrowing" the true distribution — the inverse of the "broadening" operation of section 5.1 — will yield pessimistic bounds;
- in elicitation, sometimes concern over the danger of experts' overconfidence should be replaced by concern over *under*confidence.

Our basic theorems show sufficient conditions for broader distributions to give higher reliability. But the thesis often holds even when the sufficient conditions do not hold: the correct conclusion is that since one can propagate parameter uncertainty to reliability predictions, one should do so, so that this can be taken into account in decisions.

In scenarios with broad distributions of reliability parameters, comparisons of single statistics – means or percentiles – are insufficient for choices between alternative systems. At times, a system with the worse mean value of the parameters and chosen reference percentile will give the better longer-term reliability. These scenarios also raise questions about what criteria should guide system choice or acceptance decisions (section 7.4.2), especially when choosing better reliability conflicts with uncertainty aversion.

We have mostly discussed scenarios in which the concern is safety, and thus reliability measures for functions and failures that may cause accidents; however, as pointed out in the introduction, the mathematical treatment applies as well when the concern is reliability, for instance to physical components that are meant to last for the lifetime of the systems of which they are a part.

Acknowledgements

This research was supported in part by both the Leverhulme Trust via project UnCoDe ("Uncertainty & confidence in safety arguments: effect on expert decision makers"), and the DISPO project, funded under the CINIF Nuclear Research Programme by EDF Energy Ltd., Nuclear Decommissioning Authority (Sellafield Ltd, Magnox Ltd), AWE plc and Urenco UK Ltd (the Parties). The Parties do not accept liability for any damage or loss incurred as a result of the information contained here. Views expressed are the authors' and do not necessarily represent the views of any research sponsor. Bev Littlewood, Peter Bishop and Andrey Povyakalo commented usefully on earlier drafts.

References

References

[1] Lee R. Abramson. Some drawbacks in using the mean probability. In Bilal M. Ayyub, editor, *First International Symposium on Uncertainty Modeling and Analysis*, pages 321–3, College Park, Maryland, 1990. IEEE Computer Society Press.

[2] Eugenio Alberdi, Peter Ayton, Lorenzo Strigini, and David Wright. Predictability bias: experimental results. Tech. report, CSR, City University London, 2014, in preparation.

[3] C. L. Atwood, J. L. LaChance, H. F. Martz, D. J. Anderson, M. Englehardt, D. Whitehead, and T. Wheeler. Handbook of parameter estimation for probabilistic risk assessment. Technical report, US Nuclear Regulatory Commission and Sandia National Laboratories, 2003. www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr6823.

[4] Antonia Bertolino and Lorenzo Strigini. Acceptance criteria for critical software based on testability estimates and test results. In Erwin Schoitsch, editor, *SAFECOMP 96, 15th International Conference on Computer Safety, Reliability and Security*, pages 83–94, Vienna, Austria, 1996. Springer.

[5] Antonia Bertolino and Lorenzo Strigini. Assessing the risk due to software faults: estimates of failure rate vs evidence of perfection. *Software Testing, Verification and Reliability*, 8(3):155–66, 1998.

[6] P. G. Bishop and R. E. Bloomfield. A conservative theory for long-term reliability growth prediction. *IEEE Transactions on Reliability*, 45(4):550–60, 1996.

[7] P. G. Bishop, R. E. Bloomfield, B. Littlewood, A. A. Povyakalo, and D. Wright. Toward a formalism for conservative claims about the dependability of software-based systems. *IEEE Transactions on Software Engineering*, 37(5):708–17, 2011.

[8] Peter G Bishop. Does software have to be ultra reliable in safety critical systems? In Jérémie Guiochet, editor, *SAFECOMP 2013, 32nd International Conference on Computer Safety, Reliability and Security*, Toulouse, France, 2013. Springer.

[9] R. E. Bloomfield, B. Littlewood, and D. Wright. Confidence: Its role in dependability cases for risk assessment. In *DSN 2007, 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pages 338–46, Edinburgh, U.K., 2007.

[10] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[11] C. Bunea, T. Charitos, R.M. Cooke, and G. Becker. Two-stage Bayesian models—application to ZEDB project. *Reliability Engineering and System Safety*, 90(2–3):123–30, 2005.

[12] G. de Barra. *Measure Theory and Integration.* Mathematics and Its Applications. Ellis Horwood, Chichester, 1981.

[13] M. Drouin, G. Parry, J. Lehner, G. Martinez-Guridi, J. LaChance, and T. Wheeler. Guidance on the treatment of uncertainties associated with PRAs in risk-informed decision making - main report. NUREG 1855, NRC, U.S. Nuclear Regulatory Commission, 2009.

[14] H. G. Eggleston. *Convexity.* Cambridge University Press, Cambridge, UK, 1958. Cambridge Tracts in Mathematics, Vol. 47 (1977 reprinting with corrections).

[15] EPRI. Treatment of parameter and model uncertainty for probabilistic risk assessments. Technical Report 1016737, Electric Power Research Institute (EPRI), 2008.

[16] FAA. Advisory Circular AC 25.1309-1A, Federal Aviation Administration, 1985.

[17] J. R. Fragola. Reliability and risk analysis data base development: An historical perspective. *Reliability Engineering and System Safety*, 51(2):125–36, 1996.

[18] Phyllis Frankl, Dick Hamlet, Bev Littlewood, and Lorenzo Strigini. Evaluating testing methods by delivered reliability. *IEEE Transactions on Software Engineering*, SE-24(8):586–601, August 1998.

[19] J.C. Helton and W.L. Oberkampf. Alternative representations of epistemic uncertainty. *Reliability Engineering and System Safety (special issue)*, 85(1-3):1–369, 2004.

[20] Jon C. Helton and Martin Pilch. Quantification of margins and uncertainties. *Reliability Engineering and System Safety (special issue)*, 96(9):959–1256, September 2011.

[21] Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, 1990.

[22] HSE. Step 3 C&I assessment of the EDF and AREVA UK EPR. Technical Report AR 09/038-P, Health and Safety Executive, New Reactor Build, Joint Programme Office, 2009. http://www.hse.gov.uk/newreactors/reports/step3-uk-epr-ci-assessment.pdf.

[23] D. M. Hunns and N. Wainwright. Software-based protection for Sizewell B: the regulator's perspective. *Nuclear Engineering International*, September:38–40, 1991.

[24] International Electrotechnical Commission (IEC). IEC 61508, Functional safety of electrical/ electronic/programmable electronic safety related systems.

[25] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–93, December 1906.

[26] Daniel Kahneman. *Thinking, Fast and Slow.* Farrar, Straus & Giroux, 2011.

[27] Daniel Kahneman, Paul Slovic, and Amos Tversky, editors. *Judgment under uncertainty: heuristics and biases.* Cambridge University Press, 1982.

[28] Shi-Woei Lin and Vicki M. Bier. A study of expert overconfidence. *Reliability Engineering and System Safety*, 93:711–21, 2008.

[29] J. L. Lions. Report by the inquiry board on the Ariane 5 flight 501 failure. Technical report, ESA/CNES, 19 July 1996. http://esamultimedia.esa.int/docs/esa-x-1819eng.pdf.

[30] B. Littlewood and D Wright. A Bayesian model that combines disparate evidence for the quantitative assessment of system dependability. In V Stavridou, editor, *Mathematics of Dependable Systems, II*, pages 243–58. Clarendon Press, Oxford, 1997.

[31] B. Littlewood and D. R. Wright. Reliability prediction of a software product using testing data from other products or execution environments. Tech. report no. 10 of ESPRIT DeVa project 20072, CSR, City University, London, December 1996. http://webhost.laas.fr/TSF/cabernet/deva/papers/10.ps.

[32] Bev Littlewood, Peter Popov, and Lorenzo Strigini. Modelling software design diversity - a review. *ACM Computing Surveys*, 33(2):177–208, 2001.

[33] Bev Littlewood and Lorenzo Strigini. Validation of ultra-high dependability for software-based systems. *Communications of the ACM*, 36(11):69–80, 1993.

[34] Bev Littlewood and Lorenzo Strigini. "Validation of ultra-high dependability..." — 20 years on. *Safety Systems, Newsletter of the Safety-Critical Systems Club*, May 2011. http://www.csr.city.ac.uk/people/lorenzo.strigini/ls.papers/2011_limits_20yearsOn_SCSC/.

[35] Keith W. Miller, Larry J. Morell, Robert E. Noonan, Stephen K. Park, David M. Nicol, Branson W. Murrill, and Jeffrey M. Voas. Estimating the probability of failure when testing reveals no failures. *IEEE Transactions on Software Engineering*, 18(1):33–43, 1992.

[36] P. G. Neumann. Illustrative risks to the public in the use of computer systems and related technology. http://www.csl.sri.com/users/neumann/illustrative.html.

[37] D. L. Parnas, A. J. van Schouwen, and S. P. Kwan. Evaluation of safety-critical software. *Communications of the ACM*, 33(6):636–48, 1990.

[38] Catrin Rode, Leda Cosmides, Wolfgang Hell, and John Tooby. When and why do people avoid unknown probabilities in decisions under uncertainty? testing some predictions from optimal foraging theory. *Cognition*, 72:269–304, 1999.

[39] Kizito Salako and Lorenzo Strigini. When does diversity in development reduce common failures? Insights from probabilistic modelling. *IEEE Transactions on Dependable and Secure Computing, preprints*, 2013. http://doi.ieeecomputersociety.org/10.1109/TDSC.2013.32.

[40] N. O. Siu and D. L. Kelly. Bayesian parameter estimation in probabilistic risk assessment. *Reliability Engineering and System Safety*, 62(1–2):89–116, 1998.

[41] Lorenzo Strigini and Andrey Povyakalo. Software fault-freeness and reliability predictions. In Jérémie Guiochet, editor, *SAFECOMP 2013, 32nd International Conference on Computer Safety, Reliability and Security*, Toulouse, France, 2013. Springer.

[42] J. K. Vaurio and K. E. Jänkälä. Evaluation and comparison of estimation methods for failure rates and probabilities. *Reliability Engineering and System Safety*, 91(2):209–21, 2006.

[43] M Zeelenberg. Anticipated regret, expected feedback and behavioral decision-making. *Journal of Behavioural Decision Making*, 12:93–106, 1999. Open Access publications from Tilburg University urn:nbn:nl:ui:12-80656, Tilburg University.

Appendix A Proofs about bounds

THEOREM 1: $Pr(\text{surv. } \geq t \text{ dem.}) = \int_0^1 (1-x)^t f_Q(x)\, dx \geq (1-q^*)^t$

Proof. By definition of the mean, q^*, the thesis is equivalent to:

$$\int_0^1 (1-x)^t f_Q(x)\, dx \geq \left(1 - \int_0^1 x f_Q(x)\, dx\right)^t$$

According to Jensen's inequality [25] quoted in [10, §3.1.8],[12, §6.3], given a function $l(x)$, whose 2nd derivative is non-negative for any x, we will have $E(l(x)) \geq l(E(x))$. Now, setting $l(x) = (1-x)^t$ which has the requisite property, as seen in Fig. 6 (showing the curve $l(x)$ with two cords - an alternative, equivalent[15] definition of "convexity" is that all cords are above the curve), yields the result. □

[15] In fact, in general, somewhat more inclusive, since it holds for some functions that are not everywhere twice differentiable, but equivalent for those functions that are twice differentiable.

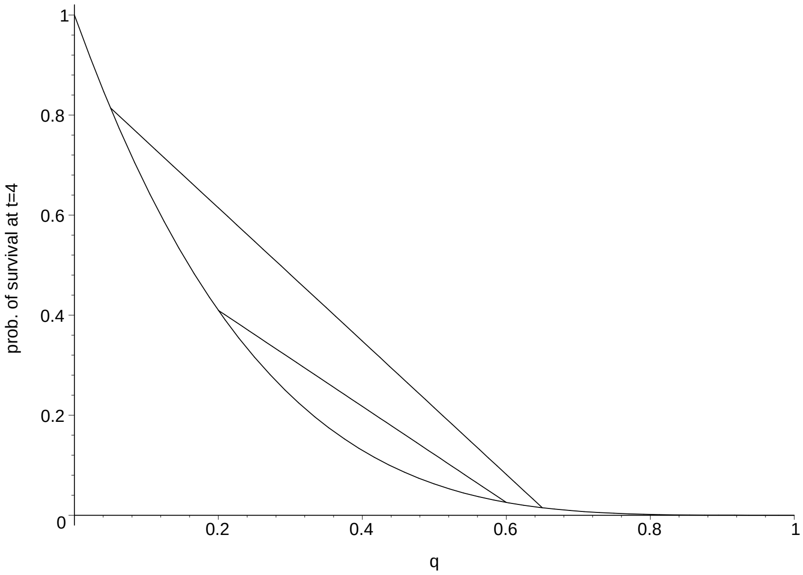


Fig. 6: Reliability $R(t) = (1-x)^t$ as a convex function of the known *pfd* x, with two of its chords shown for illustration. Geometrically a convex function is one 'bounded above by any chord, between that chord's endpoints' [14, 21, 10].

THEOREM 2: *Assuming that the distribution of the system* pfd, *Q, is*

$$\Pr(Q{=}1) = q^* , \qquad \Pr(Q{=}0) = 1{-}q^*$$

(a distribution that satisfies the condition $E(Q) = q^*$) gives an optimistic *bound on the probability of survival for at least t demands:*

$$1{-}q^* \geq \int_0^1 (1{-}x)^t f_Q(x)\, dx = \Pr(\text{survival for } {\geq} t \text{ demands})$$

Proof. Indeed, for this distribution, $\Pr(\text{surv.} \geq \text{demands}) = \Pr(Q=0)(1-0)^t + \Pr(Q=1)(1-1)^t = (1-q^*) \times 1 + q^* \times 0 = 1-q^*$. Now, for $t \geq 1$, and with the real distribution of q, $\Pr(\text{surv.} \geq t \text{ demands})$

$$= \int_0^1 (1-x)^t f_Q(x)\, dx \leq \int_0^1 (1-x) f_Q(x)\, dx = 1-q^* \qquad \square$$

Simplified linear bounds

We have the bounds of Thms. 1 & 2. A less tight, linear lower bound is found by observing that $(1-xt) \leq (1-x)^t$ (for $t \geq 1$, $0 \leq x \leq 1$). So, the set of lower and upper bounds identified can be summarised as:

$$1 - tq^* \leq (1-q^*)^t \leq \Pr(\text{survival for } {\geq} t \text{ demands}) \leq 1 - q^*.$$

Continuous Time: the upper bound on reliability is 1:

THEOREM 3. *If we consider a system with exponential reliability $e^{-\lambda t}$ conditional on a failure rate λ, and the latter is seen as a r.v. Λ with probability density function $f_\Lambda(x)$ and mean λ^*, no upper bound smaller than 1 exists for the reliability function (given the constraint $E(\Lambda) = \lambda^*$).*

Informally, the proof is by contradiction, assuming that there is such an upper bound $r < 1$, and then showing how to construct a probability distribution for Λ such that the reliability at time t is greater than r. Such a distribution is obtained by assigning a large probability mass so close to $\Lambda{=}0$ as to ensure that $R(t) > r$, and then ensuring $\mathrm{E}(\Lambda){=}\lambda^*$ by assigning the remaining probability mass to a large value $\Lambda{=}\lambda_0$ which occurs with low probability.

Proof. Suppose that, for a certain t, there is a number $r < 1$ such that for any distribution of Λ satisfying the constraint $E(\Lambda) = \lambda^*$, the reliability function $R(t) \leq r$.

Let Λ have the 2-point distribution $P(\Lambda{=}\lambda_0) = \lambda^*/\lambda_0$ with the remaining mass at 0 (which clearly satisfies $E(\Lambda) = \lambda^*$). Then, $R(t) = R(t \,|\, \Lambda = 0)P(\Lambda = 0) + R(t \,|\, \Lambda = \lambda_0)P(\Lambda = \lambda_0) = 1 \times (1 - \lambda^*/\lambda_0) + e^{-\lambda_0 t}\lambda^*/\lambda_0 > 1 - \lambda^*/\lambda_0$. Thus if we now choose $\lambda_0 > \lambda^*/(1 - r)$, we obtain $R(t){>}r$, contradicting the hypothesis. Hence, for any t, there is no upper bound $r{<}1$ for the value $R(t)$ of all reliability functions from Λ-distributions with $E(\Lambda) = \lambda^*$. □

Note: If we wish not to permit zero failure rate, we can switch to a slightly more complex proof, involving a probability mass $p_m > r$ assigned to values of $\Lambda < \lambda_m$, where $e^{-\lambda_m t} > r/p_m$.

Appendix B "Broadening" of distributions

THEOREM 4: *If a distribution of u with density function p is "broadened" according to the procedure described, and illustrated in Fig. 5, then the resulting reliability always* increases *by an amount bounded below by the positive polynomial $g(b)$ defined in (16).*

Proof. Let I and J be two finite, non-empty sub-intervals[16] of the unit interval $[0, 1]$. We write endpoints as $\underline{I} = \inf(I)$, $\overline{I} = \sup(I)$, and similarly for J. Let I be positioned entirely to the left of J, so $0 < \underline{I} \leq \overline{I} \leq \underline{J} \leq \overline{J} < 1$).

Let i and j be two non-negative functions that can only differ from zero within the intervals I and J, respectively, and have integrals between 0 and 1. Define $k > 0$ to be the ratio of their integrals so that

$$0 < \int_0^1 i(u)\,du = \int_I i(u)\,du = k\int_J j(u)\,du = k\int_0^1 j(u)\,du < 1. \tag{9}$$

Choose any positive number $b \leq \min\Big(\underline{I}, (1 - \overline{J})/k\Big)$, i.e. such that the two 'shifted intervals'[17] $I{-}b$ and $J + kb$ both lie within the unit interval. In order to retain a useful *probabilistic* interpretation we impose that the further constraint that the function p_1 given by

$$p_1(u) = p(u) + i(u + b) - i(u) - j(u) + j(u - kb) \tag{10}$$

is *non-negative* throughout the unit interval $u \in [0, 1]$.

Fig. 5 illustrates this operation performed on the distribution of u, where $1{-}u = q$ is the *pfd.* We deliberately plotted a rather implausibly complicated multi-modal (but continuous) density function p to emphasize the generality[18] of this procedure. The black graph with the discontinuities toward the LHS represents the new density p_1 obtained after applying the procedure. The grey line shows instead the original density p, assumed here (arbitrarily for the sake of this diagram) to have been a continuous function. The two subsidiary functions i, j used to perform the change are shown here as having simple, easily recognisable shapes (square and triangular, respectively), purely for the sake of this diagram's understandability. We have given density functions i and j equal 'width' and 'height' to make the ratio of their two masses easily discernable as $k = 2$. Thus the 'lighter' triangularly distributed mass is translated exactly twice the distance of the twice as 'heavy', squarely distributed mass, to ensure the preservation of the mean in the RHS of equation (11) below.

That the modification (10), of p to form p_1, is the operation described in section 5.2 follows from properties

$$\int_0^1 p_1(u)\,du = 1, \text{ and } \int_0^1 u\,p_1(u)\,du = \int_0^1 u\,p(u)\,du \tag{11}$$

[16] The reasoning we present below goes through unchanged if we drop the restriction to *intervals* and use instead arbitrary *measurable sets* with the same bounds and separating 'gaps'.

[17] Given an interval $L = [x, y]$, we define $L{+}t = [x{+}t, y{+}t]$ (generalising, for sets, to $S{+}t = \{x\,;\, x{-}t \in S\}$)

[18] In fact, we have obtained elsewhere even more general forms of the theorem about distribution broadening than this density-based view of moving fixed distributional "shapes". We address this point a little more at the end of Ap. B.

which hold since $\int_0^1 p_1(u) - p(u)\, du = \int_0^1 i(u{+}b) - i(u) - j(u) + j(u{-}kb)\, du = 0$, and $\int_0^1 u\, p_1(u)\, du - \int_0^1 u\, p(u)\, du = \int_0^1 u\, [i(u+b) - i(u) - j(u) + j(u-kb)]\, du = 0$. These latter are easy to prove, by splitting into four separate integrals and noting that the integrals of functions i and j do not change with the these translation since the functions are only non-zero in subsets of the [0,1] interval.

The property we claim for the reliability function is

$$\int_0^1 u^t\, p_1(u)\, du \geq \int_0^1 u^t\, p(u)\, du\,, \quad \text{for } t = 2, 3\ldots,$$

$$\text{i.e.,} \int_0^1 u^t\, [i(u{+}b) - i(u) - j(u) + j(u{-}kb)]\, du \geq 0\,. \tag{12}$$

In fact we can produce a non-negative t^{th} degree polynomial expression in b, and the inner endpoints of the intervals I and J, and show that this polynomial is a *lower bound for the increase* (12) in the reliability function. We use the b dependence to do this, so we regard[19] this lower-bounding expression as a polynomial function of b, and call it $g(b)$. The strategy is: to express the reliability increase (12) as the difference of values $h(b) - h(0)$ of another polynomial function $h(b)$; then to show that the derivatives of the two polynomials obey, $h'(z) \geq g'(z) \geq 0$ for $0 < z < b$. This inequality of the derivatives has the consequence that $g(b) - g(0) \leq h(b) - h(0)$, and, since it will be seen easily that in fact $g(0) = 0$, this is the sought conclusion: that $g(b)$ must be a non-negative lower bound for (12).

So we first define the polynomial $h(z)$, of degree t, by

$$h(z) = \int u^t\, [j(u{-}kz) + i(u{+}z)]\, du = \int (u{+}kz)^t\, j(u) + (u{-}z)^t\, i(u)\, du \tag{13}$$

where the integrals are over the whole real line (given that the integrand is zero outside the above defined finite intervals I and J of support), and where we have used simple translation changes of variable to obtain the second form. Differentiating this polynomial gives a degree $t{-}1$ polynomial $h'(z)$ that can be compactly represented by differentiating (13) under the integral to produce

$$h'(z) = t \int k(u + kz)^{t-1}\, j(u) - (u - z)^{t-1}\, i(u)\, du\,. \tag{14}$$

Our lower bound on this derivative polynomial is

$$\begin{aligned} h'(z) \quad &\geq \quad t \int k(\underline{J} + kz)^{t-1}\, j(u) - (\overline{I} - z)^{t-1}\, i(u)\, du \\ &= \quad tk \left[(\underline{J} + kz)^{t-1} - (\overline{I} - z)^{t-1} \right] \int j(u)\, du \end{aligned} \tag{15}$$

by (9). As explained above, we call the latter $g'(z)$, being the derivative of a polynomial $g(z) = \left[(\underline{J}{+}kz)^t - \underline{J}^t + k(\overline{I}{-}z)^t - k\overline{I}^t \right] \int j(u)\, du$. It is self-evident from the definitions that $g(0) = 0$ and $g'(z) \geq 0$ throughout $0 \leq z \leq b$. (We remark that $g(z)$ is merely $h(z)$ with, in its coefficients, powers of the interval endpoints replacing the corresponding i and j moments mentioned above.) So we have done all that was required to conclude that the increase in reliability, resulting from the "movement of mass" described, satisfies

$$\begin{aligned} \int_0^1 u^t p_1(u)\, du - \int_0^1 u^t p(u)\, du &= \int_0^1 u^t\, [i(u{+}b) - i(u) - j(u) + j(u{-}kb)]\, du \\ &= h(b) - h(0) \;\geq\; g(b) \\ &= \left[(\underline{J}{+}kb)^t - \underline{J}^t + k(\overline{I}{-}b)^t - k\overline{I}^t \right] \int_0^1 j(u)\, du > 0. \end{aligned} \tag{16}$$

□

[19] I.e., we imagine "collecting powers of b" so that the other quantities – k and the inner I, J endpoints, on which the lower bound $g(b)$ depends – are incorporated into the coefficients of b powers.

Generalisations Although this "mass-shifting" procedure and associated proofs have been presented for intervals $I{-}b$, I, J, & $J{+}kb$ (with $b > 0$ and k related to the two densities i, j as described above), nevertheless, they generalise to arbitrary measurable sets in place of intervals, and to arbitrary positive, finite *measures* [12] on those two sets. That the procedure increases reliability does not rely in any way on the *existence* of probability density, provided that

- the 'leftmost pair' of measures are related by simple translation; and similarly the 'rightmost pair'
- the two "translation distances" respectively relating the leftmost and the rightmost pairs are in inverse proportion to the sizes of the two measures. Note that:
  - The leftmost pair of support sets may intersect, or not. Similarly the rightmost pair. We need only each pair to be related by the specified translation
  - The supremum of the righthand member of the leftmost pair of sets must not exceed the infimum of the lefthand member of the rightmost pair of sets. In other words, the two 'inner sets' of the foursome must *not intersect*, and in fact must lie on opposite sides of some dividing point (they may touch at the point)
- the two measures corresponding here to densities i and j must be positive, finite and with supports within the specified "inner pair" of the four sets.

In this way, the results extend to measures which may not possess density functions (with respect to Lebesgue measure). The argument about the total measure and its first moment being unaffected by the procedure goes through essentially as above. Similarly the argument for the increase in the expected value of the function $u \mapsto u^t$ where $t = 2, 3, \ldots$ also goes through (on the nonnegative real half-line). In fact it goes through for any sufficiently well-behaved convex function on real $t{\geq}1$ (though h and h' are then no longer necessarily polynomials).

It is possible to extend further, using measures expressed in terms of their cumulative distribution functions (c.d.f.s). The "broadening" procedure (and associated proof) for this extension is very general and uses five *arbitrary*, finite, positive *measures* in place of the five densities $(p{-}i{-}j)(u)$, $i(u)$, $j(u)$, $i(u{+}b)$, $j(u{-}kb)$ in the above proof. Using this c.d.f.-based representation, the measures corresponding to densities $i(u)$ and $i(u{+}b)$ in the above proof are then *no longer required* to be related by simple *translation*. Informally, this allows the distributional "shapes" of mass to change[20] as that mass is shifted. The generalisation uses an equivalent condition on the means, keeping the fixed-$E(u)$ constraint. Of course, if not simple translation of measures then *some* replacement which generalizes this idea is still needed to characterize "mass moving *apart"*. To do this using c.d.f.s is in fact rather simple: we merely require that the two c.d.f. functions (counterparts of the densities $i(u)$ and $i(u{+}b)$ above) are stochastically ordered (in comparison to one another) in the simplest pointwise sense; and similarly the other two c.d.f.s (replacing what, above, were densities $j(u)$ and $j(u{-}kb)$). We omit the detail of proof this extension of Thm. 4 which employs integration by parts for Stieltjes-type integrals.

Appendix C Multiple components or parameters

THEOREM 5: *The reliability function of a* series *system made up of two components that fail independently is neither convex nor concave on any subset of the unit square.*

Proof. The probability of survival of this "series" system for t demands, given specific *pfd* values for the two components q_A and q_B, is given by equation (8):

$$\Pr(\text{survival for } {\geq}t \text{ demands}; q_A, q_b) = (1 - q_A)^t (1 - q_B)^t$$

[20] even to the extent that points of concentrated mass may become continuously dispersed, or vice versa

and thus when averaging over the uncertainty the corresponding probability becomes $\int_0^1\int_0^1 (1-q_A)^t(1-q_B)^t f_{Q_A,Q_B}(q_A,q_B)\,dq_A dq_B$.
Were the parameterised reliability function (8) convex over the unit square, then the 2 dimensional version of Jensen's inequality would be applicable just as in the univariate case examined above. However, now this is emphatically not the case. Examining the Hessian matrix of this function (8), one finds it to be a non-negative scalar multiple of the matrix

$$H = \begin{bmatrix} (t-1)(1-q_B)^2 & t(1-q_B)(1-q_A) \\ t(1-q_B)(1-q_A) & (t-1)(1-q_A)^2 \end{bmatrix}$$

of $\det(H) = -(2t-1)(1-q_A)^2(1-q_B)^2$ is everywhere *negative* (since $t = 1, 2, \ldots$), while its diagonal terms are everywhere *non*-negative. It follows that for no t is (8) either convex or concave on any open subset of the unit square. (See e.g. [14, p51], [21, §7.2.5], [10, §3.1.4][21].) □

It follows that, for each $t = 2, 3, \ldots$, there exist (without making any assertion about their realism) bivariate distributions of (Q_A, Q_B) such that Pr(surv. $\geq t$ dem.) $> (1-q_A^*)^t(1-q_B^*)^t$, and, for that same t, there exist other distributions for which this inequality is reversed.

THEOREM 6: *The system reliability function* $(1-q_Aq_B)^t$ *of a* parallel *system made up of two components that fail independently is concave on no subset of the unit square, and convex on a convex subset* F *of the unit square precisely if* F *is a subset of the region* D *on which this reliability* $(1-q_Aq_B)^t$ *is below* $(1-\frac{1}{2t-1})^t$ *in value.*

Proof. The system *pfd*, $q_{AB} = q_Aq_B$, when averaged over our uncertainty gives Pr(surv. $\geq t$ dem.) $= \int_0^1\int_0^1 (1-q_Aq_B)^t f_{Q_A,Q_B}(q_A,q_B)\,dq_A dq_B$. Now, the Hessian matrix of $(1-q_Aq_B)^t$ is a non-negative scalar multiple of $H = \begin{bmatrix} (t-1)q_B^2 & tq_Aq_B - 1 \\ tq_Aq_B - 1 & (t-1)q_A^2 \end{bmatrix}$ with again – as for the *series* case – non-negative diagonal elements. However, the determinant $\det(H) = (1-q_Aq_B)\left[(2t-1)q_Aq_B - 1\right]$ is interesting in two respects:

1. given the non-negativity of the diagonals of matrix H, we can conclude that, for any fixed t, on any convex, open subset F of the unit square, our survival probability $(1-q_Aq_B)^t$ is convex on the set F if and only if this determinant is non-negative throughout F. Otherwise is it non-convex, and it can never be concave on F.

2. both the survival probability $(1-q_Aq_B)^t$ whose convexity is under investigation, and the sign of its Hessian determinant are fully determined by product q_Aq_B.

It follows from this second observation that Jensen's inequality will hold over such a set F precisely if F is a subset of the region D of the unit square on which the reliability function $(1-q_Aq_B)^t$ is below $(1-\frac{1}{2t-1})^t$. For D is precisely the region on which the product q_Aq_B exceeds $\frac{1}{2t-1}$, i.e. $\det(H) \geq 0$. Thus, the bounds based on Jensen's inequality extend to parallel systems, allowing one to state that Pr(surv. $\geq t$ demands) $= \int_0^1\int_0^1 (1-q_Aq_B)^t f_{Q_A,Q_B}(q_A,q_B)\,dq_A dq_B \geq (1-q_A^*q_B^*)^t$ only in scenarios that turn out to be unlikely to present themselves in concrete applications: the bound applies if we consider it impossible (i.e. assign prior probability zero) that the pair q_A, q_B could belong to any subset of the unit square which contains even a single point at which the survival probability $(1-q_Aq_B)^t$ would exceed $(1-\frac{1}{2t-1})^t$. Given that the sequence $\left\langle \left(1-\frac{1}{2t-1}\right)^t \right\rangle_{t=1}^{\infty}$ is increasing to limit $\exp(-0.5) = 0.6065$, this can be restated as: "we can obtain our Jensen-based lower bound by plugging in the means q_A^*, q_B^* if we are certain that the best system we can possibly have built has less probability of fulfilling the mission assigned

[21] Unlike other references, Boyd [10] omits – we suspect erroneously – to require as a premise of his statement *continuous* 2$^{\text{nd}}$-order partial derivatives.

than some figure, depending on assigned lifetime (or mission) length t, which, for any t, will always be under 61%." Hence, in the parallel case, it is not that the Jensen-based worst case bound is never applicable; but rather that it can be applicable only for systems with comparatively low reliabilities (not the scenario of interest in this paper). □